Propensity Score Methods for Merging Observational and Experimental Datasets

Evan Rosenman, Michael Baiocchi, and Art B. Owen

Stanford University

April 2018

Contents

Abstract

We consider merging information from a randomized controlled trial (RCT) into a much larger observational database (ODB), for the purpose of estimating a treatment effect. In our motivating setting, the ODB has better representativeness (external validity) while the RCT has genuine randomization. We work with strata defined by propensity score in the ODB. For all subjects in the RCT, we find the propensity score they would have had, had they been in the ODB. We develop and simulate two hybrid methods. The first method simply spikes the RCT data into their corresponding ODB stratum. The second method takes a data driven convex combination of the ODB and RCT treatment effect estimates within each stratum. We develop delta method estimates of the bias and variance of these methods and we simulate them. The spike-in method works best when the RCT covariates are drawn from the same distribution as in the ODB. When the RCT inclusion criteria are very different than those of the ODB, then the spiked-in estimate can be severely biased and the second, dynamic, method works better.

1 Introduction

The increasing availability of large, observational datasets poses opportunities and challenges for statistical methodologists. These datasets often "contain detailed information about the target population of interest," meaning they could have great utility for estimating the causal effects of a proposed intervention, such as a public health initiative (Hartman et al., 2013). Yet assignment to the test or control group is almost never random in these data, making standard methods prone to misstate the effect of the intervention.

Randomized control trials (RCTs) make causal inference substantially easier, because the researcher controls assignment of the intervention. Yet RCTs present their own challenges. A commonly raised concern is that "estimates from RCTs $\cdots$ may lack external validity" (Hartman et al., 2013) due to the sampling scheme used to enroll participants. Moreover, in cases where the treatment effect varies by subpopulation, "experiments have to be very large and, in general, prohibitively costly" whereas "observational data is often available in much larger quantities" (Peysakhovich and Lada, 2016).

These issues motivate a hybrid approach, which makes use of the availability, size, and representativeness of observational data as well as the randomization inherent in RCT data. The fundamental tool used in our approach is the propensity score: the estimated conditional probability of exposure to the intervention, given observed covariates.

Rosenbaum and Rubin (1984) showed that comparing treated individuals against control individuals for whom the propensity score is approximately equal yields a substantial reduction in bias. The propensity score is widely used in analysis of observational datasets, as comparing test and control units with similar propensity scores "tends to balance observed covariates that were used to construct the score" (Joffe and Rosenbaum, 1999). It has been less widely used in the context of RCTs. But propensity-based methods "have been shown to be useful even in randomized control settings, where the assignment mechanism is known and independent of the covariates" (Xu and Kalbfleisch, 2010), because they correct for chance imbalance in covariates as well as biased designs.

In using the propensity score, we make a simplifying assumption: we allow for a heterogeneous treatment effect but assume that it varies only as a function of the propensity. The treatment effect for subjects in the RCT is assumed to be the same function of the propensity that holds in the ODB. The distribution of covariates in the RCT can however be much different from the ODB due to factors such as varying enrollment criteria.

We assume that the covariate distribution in the ODB is the same as that of the target population. We use data from the RCT to provide much-needed control subjects in the strata with a high propensity in the ODB as well as test group subjects in the strata with a low propensity in the ODB.

Our two main estimators are a spiked-in estimator that simply merges the ODB and RCT within each stratum and a "dynamically weighted" estimator that mimics the best possible convex combination of estimates from the two populations, within each stratum.

The remainder of the paper is organized as follows. In Section 2, we define our notation, assumptions, estimand and estimators, including the spiked-in and dynamic weighting estimators that we propose. We work in the potential outcomes framework, in which treatment effects are ratio estimators due to the random numbers of subjects in each condition. For the large sample sizes of interest to us, delta method approximations to the mean and variance are accurate enough. Section 3 presents delta method estimates of the within-stratum bias and variance for our estimators. There we see theoretically that the spiked-in estimator can have an enormous bias if the covariates in the RCT do not follow the same distribution as those of the ODB. Section 4 gives some numerical illustrations of our method for an ODB of size 5,000 and an RCT of size 200. When the RCT covariates follow the ODB's distribution, then the spiked-in estimator brings a large reduction in mean squared error over the ODB-only estimate. If, however, enrollment criteria bias the RCT, then the spiked-in estimator can be much worse than the ODB-only estimate. In either case, the dynamic weighted estimator brings an improvement over the ODB-only estimate. Section 5 summarizes our conclusions and an Appendix contains two of the lengthier proofs.

2 Notation, assumptions and estimators

Some subjects belong to the randomized controlled trial (RCT) and others to the observational database (ODB). We assume that no subject is in both data sets. We write $i \in \mathcal{R}$ if subject i is in the RCT and $i \in \mathcal{O}$ otherwise. Subject i has an outcome $Y_i \in \mathbb{R}$ and some covariates that we encode in the vector $\boldsymbol{x}_i \in \mathbb{R}^d$. Subject i receives either the test condition or the control condition.

The condition of subject i is given by a treatment variable $W_i \in \{0, 1\}$ where

$W_i = 1$ if subject i is in the test condition (and 0 otherwise). Some formulas simplify when we can use parallel notation for both test and control settings. Accordingly we introduce $W_{it} = W_i$ and $W_{ic} = 1 - W_i$. Other formulas look better when focused on the test condition. For instance, letting $p_{it} = \Pr(W_{it} = 1)$ and $p_{ic} = \Pr(W_{ic} = 1)$, the expression $p_{it}(1-p_{it})$ is immediately recognizable as a Bernoulli variance and is preferred to $p_{it}p_{ic}$.

2.1 Model

We adopt the potential outcomes framework of Neyman and Rubin. See Rubin (1974). Subject i has two potential outcomes, Y_{it} and Y_{ic}, corresponding to test and control conditions respectively. Then $Y_i = W_{it}Y_{it} + W_{ic}Y_{ic}$. The potential outcomes (Y_{it}, Y_{ic}) are non-random and we will assume that they are bounded. We work conditionally on the observed values of covariates and so $\boldsymbol{x}_i$ are also non-random.

All of the randomness comes from the treatment variables W_i. We use the notation $\mathrm{Bern}(p)$ for Bernoulli random variables taking the value 1 with probability p and 0 with probability $1 - p$. The ODB and RCT differ in how the W_i are distributed.

Assumption 1 (ODB sampling)**.** If $i \in \mathcal{O}$, then $W_i \sim \mathrm{Bern}(p_i)$ independently where $p_i = e(\boldsymbol{x}_i)$ with $0 < p_i < 1$.

The function $e(\cdot)$ in Assumption 1 is a propensity. Because the propensity depends only on $\boldsymbol{x}$, and is never 0 or 1, the ODB has a strongly ignorable treatment assignment (Rosenbaum and Rubin, 1984). Because the W_i are independent, the outcome for subject i is unaffected by the treatment $W_{i'}$ for any subject $i' \neq i$. That is, our model for the ODB satisfies the stable unit treatment value assumption or SUTVA (Imbens and Rubin, 2015).

Assumption 2 (RCT sampling)**.** If $i \in \mathcal{R}$, then $W_i \sim \mathrm{Bern}(p_r)$ independently for a common probability $0 < p_r < 1$.

The RCT will commonly have $p_r = 1/2$ but we do not assume this. Our RCT model also has a strongly ignorable treatment assignment and it too satisfies the SUTVA. We additionally assume that the ODB is independent of the RCT.

2.2 Stratification

Our comparison of treatment versus control is based on stratification by propensity as described by Imbens and Rubin (2015). This is one of several matching strategies mentioned in Stuart and Rubin (2007).

We use K strata defined by propensity intervals. For the ODB these are

$$\mathcal{O}_k = \Bigl\{ i \in \mathcal{O} \bigm| \frac{k-1}{K} < e(\boldsymbol{x}_i) \leqslant \frac{k}{K} \Bigr\}, \quad k = 1, \dots, K.$$

The RCT is similarly stratified via

$$\mathcal{R}_k = \Bigl\{ i \in \mathcal{R} \bigm| \frac{k-1}{K} < e(\boldsymbol{x}_i) \leqslant \frac{k}{K} \Bigr\}, \quad k = 1, \dots, K.$$

Symbols	Meaning
i, k, ω_k	Subject and stratum indices, stratum weights
$\mathcal{O}$, $\mathcal{R}$, $\mathcal{O}_k$, $\mathcal{R}_k$	ODB and RCT subject sets and strata
$\boldsymbol{x}_i$, $e(\boldsymbol{x}_i)$	Covariates and ODB propensities
Y_{it}, Y_{ic}, Y_i	Test, control and observed responses
W_{it}, W_{ic}, W_i	Test, control and observed indicators. $W_i \equiv W_{it}$
$\tau_{ok} = \tau_{rk} = \tau_k$	Stratum treatment effects: ODB, RCT, merged
n_{ok}, n_{rk}, n_k	Stratum sample sizes: ODB, RCT, merged
$\hat{\tau}_{ok}$, $\hat{\tau}_{rk}$	ODB and RCT estimates of $\hat{\tau}_k$
$\hat{\tau}_{sk}$, $\hat{\tau}_{wk}$, $\hat{\tau}_{dk}$	Spiked, weighted, dynamic estimates of $\hat{\tau}_k$
μ_{okt}, μ_{okc}	Average potential responses by ODB stratum
μ_{rkt}, μ_{rkc}	Average potential responses by RCT stratum
p_{okt}, p_{rkt}	Average propensities by ODB and RCT strata
p_{rkc}, p_{rkc}	One minus average propensities
s_{okt}, s_{okc}	Response-propensity covariances in the ODB

Table 1: Summary of notation used.

Note that the RCT is stratified according to the propensity that those observations would have had, had they been in the ODB. Imbens and Rubin (2015) suggest strata containing approximately equal numbers of observations. We have instead given them equal sized propensity ranges. In practice, some small strata might have to be merged together. Sometimes we refer to strata as 'bins'.

We work here as though the propensity function e is known exactly. In practice, e will be replaced by an estimated propensity fit to the ODB. We suppose that the ODB is large enough to obtain a good propensity estimate. Under Assumption 1, the true propensity is a function of the observed variables $\boldsymbol{x}_i$.

The sample sizes of the ODB and RCT are n_o and n_r respectively. Ordinarily $n_o \gg n_r$. The ODB and RCT sample sizes within stratum k are n_{ok} and n_{rk}. The within-stratum average treatment effects are

$$\tau_{ok} = \frac{1}{n_{ok}} \sum_{i \in \mathcal{O}_k} Y_{it} - Y_{ic} \quad \text{and} \quad \tau_{rk} = \frac{1}{n_{rk}} \sum_{i \in \mathcal{R}_k} Y_{it} - Y_{ic}, \tag{1}$$

where means over empty strata are taken to be 0 in (1).

Assumption 3. For $k = 1, \dots, K$, if $n_{ok} > 0$ and $n_{rk} > 0$ then $\tau_{ok} = \tau_{rk}$ and we call their common value τ_k.

Assumption 3 leaves τ_k undefined when $\min(n_{ok}, n_{rk}) = 0$. If only one of n_{ok} and n_{rk} is positive then we take its treatment effect for τ_k. If both are 0, then we will not need τ_k.

Assumption 4. For all $i \in \mathcal{O}_k \cup \mathcal{R}_k$, $Y_{it} - Y_{ic} = \tau_k$.

Assumption 4 is an idealization that simplifies some derivations, and we need it in one instance to estimate a quantity that depends on both potential outcomes of a single subject. In some of our simulations we will relax that assumption to make $Y_{it} - Y_{ic}$ constant for all units i at a fixed propensity score, rather than within a stratum. Xie et al. (2012) have argued for analyzing the pattern of treatment effects solely as a function of the propensity score, an approach taken by a number of social science researchers (Brand and Davis, 2011; DellaPosta, 2013). Because our strata are based on propensity, Assumption 4 is very nearly true under the model of Xie et al. (2012).

Assumption 4 could be made more realistic by stratifying on both the propensity score and a ‘prognostic score’ predicting the potential outcome without treatment. Many of our results do not require strata to be constructed exclusively by propensity scores. Our simulations and discussion do focus exclusively on stratification by propensity score.

2.3 Estimators

Our estimand is a global average treatment effect defined by

$$\tau = \sum_{k=1}^{K} \omega_k \tau_k$$

for weights $\omega_k \geqslant 0$ with $\sum_{k=1}^{K} \omega_k = 1$. The weights can be chosen to match population characteristics. Our choice is to take $\omega_k = n_{ok}/n_o$ which is reasonable when the ODB represents the target population of interest. With this choice, $\omega_k = 0$ whenever $n_{ok} = 0$ and we have a well defined τ_k for every stratum that contributes to τ. We may still have $n_{rk} = 0$ for some strata with $\omega_k > 0$.

We now introduce some estimators designed to make use of the advantages of both the RCT and the ODB data. Our estimators all take the form $\sum_k \omega_k \hat{\tau}_k$ for different within-stratum estimates $\hat{\tau}_k$.

Our two simplest proposed estimators each use just one of the two populations. The ODB-only estimate of the treatment effect in stratum k is

$$\hat{\tau}_{ok} = \frac{\sum_{i \in \mathcal{O}_k} W_{it} Y_{it}}{\sum_{i \in \mathcal{O}_k} W_{it}} - \frac{\sum_{i \in \mathcal{O}_k} W_{ic} Y_{ic}}{\sum_{i \in \mathcal{O}_k} W_{ic}}. \tag{2}$$

Then $\hat{\tau}_o = \sum_k \omega_k \hat{\tau}_{ok}$. A potential problem with $\hat{\tau}_o$ is that small values of k, corresponding to the left-most bins, have subjects with small propensity values. Then $\mathcal{O}_k$ may contain very few observations with $W_{it} = 1$ and $\hat{\tau}_{ok}$ may have high variance. Similarly for large k, $\mathcal{O}_k$ may contain very few observations with $W_{ic} = 1$ which again leads to high variance. That is, the ‘edge bins’ can have very skewed sample sizes causing problems for $\hat{\tau}_o$.

The ODB estimate (2) is a difference of ratio estimators, because the denominators are random. We will see in Section 3 that there can also be a severe bias in the edge bins.

An analogous RCT-only estimator is $\hat{\tau}_r = \sum_k \omega_k \hat{\tau}_{rk}$ where

$$\hat{\tau}_{rk} = \frac{\sum_{i \in \mathcal{R}_k} W_{it} Y_{it}}{\sum_{i \in \mathcal{R}_k} W_{it}} - \frac{\sum_{i \in \mathcal{R}_k} W_{ic} Y_{ic}}{\sum_{i \in \mathcal{R}} W_{ic}}. \tag{3}$$

Because the RCT assigns treatments with constant probability, the edge bins have less skewed treatment outcomes. However, because the RCT is small, we may find that several of the strata have very small sample sizes n_{rk}.

Our first hybrid estimator is $\hat{\tau}_s = \sum_k \omega_k \hat{\tau}_{sk}$, where

$$\hat{\tau}_{sk} = \frac{\sum_{i\in\mathcal{O}_k} W_{it}Y_{it} + \sum_{i\in\mathcal{R}_k} W_{it}Y_{it}}{\sum_{i\in\mathcal{O}_k} W_{it} + \sum_{i\in\mathcal{R}_k} W_{it}} - \frac{\sum_{i\in\mathcal{O}_k} W_{ic}Y_{ic} + \sum_{i\in\mathcal{R}_k} W_{ic}Y_{ic}}{\sum_{i\in\mathcal{O}_k} W_{ic} + \sum_{i\in\mathcal{R}_k} W_{ic}}. \tag{4}$$

The RCT data are ‘spiked’ into the ODB strata. This spiked-in estimator can improve upon the ODB estimator by increasing the number of treated units in the low propensity edge bins and increasing the number of control units in the high propensity edge bins. Even a small number of such balancing observations can be extremely valuable.

The spiked-in estimator is not a convex combination of $\hat{\tau}_{ok}$ and $\hat{\tau}_{rk}$, because the pooling is first done among the test and control units. Our final two estimators *are* constructed as convex combinations of $\hat{\tau}_{ok}$ and $\hat{\tau}_{rk}$.

The weighted average estimator $\hat{\tau}_w$ uses

$$\hat{\tau}_{wk} = \lambda_k \hat{\tau}_{ok} + (1-\lambda_k)\hat{\tau}_{rk}, \quad \text{where} \quad \lambda_k = \frac{n_{ok}}{n_{ok}+n_{rk}}. \tag{5}$$

It weights $\hat{\tau}_{rk}$ and $\hat{\tau}_{ok}$ according to the number of data points involved in each estimate.

Our final estimator is a “dynamic weighted average” $\hat{\tau}_d$. It uses weights for $\hat{\tau}_{rk}$ and $\hat{\tau}_{ok}$ that are estimated from the data. Those weights are chosen to minimize an estimate of mean squared error (MSE) derived using the delta method in the following section. While the precise form of this estimator will be discussed next, we can observe its approximate optimality via the following result, recalling that the RCT estimator will in general be unbiased.

Proposition 1. *Let $\hat{\phi}_1$ and $\hat{\phi}_2$ be independent estimators of a common quantity ϕ, with bias, variance and mean squared errors,* $\mathrm{Bias}(\hat{\phi}_1) \in (-\infty,\infty)$, $\mathrm{Bias}(\hat{\phi}_2) = 0$, $\mathrm{var}(\hat{\phi}_j)$ *and* $\mathrm{MSE}(\hat{\phi}_j) \in (0,\infty)$ *for $j = 1, 2$. For $c \in \mathbb{R}$, let $\hat{\phi}_c = c\hat{\phi}_1 + (1-c)\hat{\phi}_2$. Then*

$$c_* \equiv \operatorname*{argmin}_c \mathrm{MSE}(\hat{\phi}_c) = \frac{\mathrm{var}(\hat{\phi}_2)}{\mathrm{MSE}(\hat{\phi}_1) + \mathrm{var}(\hat{\phi}_2)}.$$

This linear combination has

$$\begin{aligned}
\mathrm{Bias}(\hat{\phi}_{c_*}) &= \frac{\mathrm{Bias}(\hat{\phi}_1)\mathrm{MSE}(\hat{\phi}_2)}{\mathrm{MSE}(\hat{\phi}_1)+\mathrm{MSE}(\hat{\phi}_2)},\\
\mathrm{var}(\hat{\phi}_{c_*}) &= c_*^2\mathrm{var}(\hat{\phi}_1) + (1-c_*)^2\mathrm{var}(\hat{\phi}_2), \quad \textit{and}\\
\mathrm{MSE}(\hat{\phi}_{c_*}) &= \frac{\mathrm{MSE}(\hat{\phi}_1)\mathrm{var}(\hat{\phi}_2)}{\mathrm{MSE}(\hat{\phi}_1)+\mathrm{var}(\hat{\phi}_2)}.
\end{aligned} \tag{6}$$

Proof. Independence of the $\hat{\phi}_j$ yields $\mathrm{var}(\hat{\phi}_c) = c^2\mathrm{var}(\hat{\phi}_1) + (1-c)^2\mathrm{var}(\hat{\phi}_2)$ while linearity of expectation yields $\mathrm{Bias}(\hat{\phi}_c) = c\mathrm{Bias}(\hat{\theta}_1)$. Optimizing $\mathrm{MSE}(\hat{\phi}_c)$ over c yields the result. □

3 Delta method results

In this section we develop some delta method moment approximations. Let $\boldsymbol{X}$ be a random vector with mean μ and a finite covariance matrix. Let f be a function of $\boldsymbol{X}$ that is twice differentiable in an open set containing μ and let f_1 and f_2 be first and second order Taylor approximations to f around μ. Then the delta method mean and variance of $f(\boldsymbol{X})$ are

$$\mathbb{E}_\delta(f(\boldsymbol{X})) = \mathbb{E}(f_2(\boldsymbol{X})) \quad \text{and} \quad \mathrm{var}_\delta(f(\boldsymbol{X})) = \mathrm{var}(f_1(\boldsymbol{X}))$$

respectively.

Sometimes, to combine estimates, we will need a delta method mean for a weighted sum of those estimates. We will also need a delta method variance for a weighted sum of independent random variables. We use the following natural expressions

$$\mathbb{E}_\delta\biggl(\sum_j \lambda_j \hat\tau_j\biggr) = \sum_j \lambda_j \mathbb{E}_\delta(\hat\tau_j) \tag{7}$$

$$\mathrm{var}_\delta\biggl(\sum_j \lambda_j \hat\tau_j\biggr) = \sum_j \lambda_j^2 \mathrm{var}_\delta(\hat\tau_j), \quad \text{for independent } \hat\tau_j \tag{8}$$

without making recourse to Taylor approximations.

3.1 Population quantities

We will study our estimators in terms of some population quantities. These involve some unobserved values of Y_{it} or Y_{ic}. For instance, the test and control stratum averages in the ODB are

$$\mu_{okt} = \frac{\sum_{i\in\mathcal{O}_k} Y_{it}}{n_{ok}} \quad \text{and} \quad \mu_{okc} = \frac{\sum_{i\in\mathcal{O}_k} Y_{ic}}{n_{ok}}$$

and it is typical that both of these are unobserved. Corresponding values for the RCT are μ_{rkt} and μ_{rkc}.

When we merge ODB and RCT strata we will have to consider a kind of skew in which the within-stratum mean responses above differ between the two data sets. To this end, define

$$\Delta_{kt} = \mu_{okt} - \mu_{rkt} \quad \text{and} \quad \Delta_{kc} = \mu_{okc} - \mu_{rkc}.$$

Under Assumption 3, $\Delta_{kt} = (\tau_k + \mu_{okc}) - (\tau_k + \mu_{rkc}) = \Delta_{kc}$. We will use $\Delta_k = \Delta_{kt} = \Delta_{kc}$.

Now we define several other population quantities. Let $\mathcal{S}$ be a finite non-empty set of $n = n(\mathcal{S})$ indices such as one of our strata $\mathcal{O}_k$ or $\mathcal{R}_k$. For each $i \in \mathcal{S}$, let $(Y_{it}, Y_{ic}) \in [-B, B]^2$ be a pair of bounded potential outcomes and let $W_i = W_{it}$ be independent $\mathrm{Bern}(p_i)$ random variables and let $W_{ic} = 1 - W_{it}$. Some of our results add the condition that all $p_i \in [\epsilon, 1-\epsilon]$ for some $\epsilon > 0$.

For $\mathcal{S}$ so equipped, we define average responses

$$\mu_t = \mu_t(\mathcal{S}) = \frac{1}{n}\sum_{i\in\mathcal{S}} Y_{it} \quad \text{and} \quad \mu_c = \mu_c(\mathcal{S}) = \frac{1}{n}\sum_{i\in\mathcal{S}} Y_{ic}. \tag{9}$$

For example, μ_{okt} above is $\mu_t(\mathcal{O}_k)$. We use average treatment probabilities

$$p_t = p_t(\mathcal{S}) = \frac{1}{n}\sum_{i\in\mathcal{S}} p_i \quad \text{and} \quad p_c = p_c(\mathcal{S}) = 1 - p_t(\mathcal{S}). \tag{10}$$

These become p_{okt}, p_{okc}, p_{rkt} and p_{rkc} in a natural notation when $\mathcal{S}$ is $\mathcal{O}_k$ or $\mathcal{R}_k$.

The above quantities are averages over i uniformly distributed in $\mathcal{S}$ as distinct from expectations with respect to random W_i. We also need some covariances of this type between response and propensity values,

$$\begin{aligned} s_t = s_t(\mathcal{S}) &= \frac{1}{n}\sum_{i\in\mathcal{S}} Y_{it}p_i - \mu_t p_t \quad \text{and} \\ s_c = s_c(\mathcal{S}) &= \frac{1}{n}\sum_{i\in\mathcal{S}} Y_{ic}(1-p_i) - \mu_c p_c. \end{aligned} \tag{11}$$

We will find that these quantities play an important role in bias. If for instance the larger values of Y_{it} tend to co-occur with higher propensities p_i then averages are biased up.

The delta method variances of our estimators depend on the following weighted averages of squares and cross products

$$\begin{aligned} S_{tt} = S_{tt}(\mathcal{S}) &= \frac{1}{n}\sum_{i\in\mathcal{S}} p_i(1-p_i)(Y_{it}-\rho_t)^2, \\ S_{cc} = S_{cc}(\mathcal{S}) &= \frac{1}{n}\sum_{i\in\mathcal{S}} p_i(1-p_i)(Y_{ic}-\rho_c)^2, \quad \text{and} \\ S_{tc} = S_{tc}(\mathcal{S}) &= \frac{1}{n}\sum_{i\in\mathcal{S}} p_i(1-p_i)(Y_{it}-\rho_t)(Y_{ic}-\rho_c), \end{aligned} \tag{12}$$

where $\rho_t = \rho_t(\mathcal{S}) = \mu_t(\mathcal{S}) + s_t(\mathcal{S})/p_t(\mathcal{S})$ and $\rho_c = \rho_c(\mathcal{S}) = \mu_c(\mathcal{S}) + s_c(\mathcal{S})/p_c(\mathcal{S})$.

Proposition 2. *Let $\mathcal{S}$ be $\mathcal{O}_k$, $\mathcal{R}_k$ or $\mathcal{O}_k \cup \mathcal{R}_k$. Then under Assumption 4, $s_c(\mathcal{S}) = -s_t(\mathcal{S})$.*

Proof. Under Assumption 4, we can set $Y_{it} = Y_{ic} + \tau_k$ and $\mu_t = \mu_c + \tau_k$ in (11). □

3.2 Main theorem

We will compare the efficiency of our five estimators using their delta method approximations. We state two elementary propositions without proof and then give our main theorem. Results for our various estimators are mostly direct corollaries of that theorem.

Proposition 3. *Let x and y be jointly distributed random variables with means $x_0 \neq 0$ and y_0 respectively, and finite variances. Let $\rho = y_0/x_0$. Then*

$$\mathbb{E}_\delta\Big(\frac{y}{x}\Big) = \rho - \frac{\operatorname{cov}(y-\rho x, x)}{x_0^2}, \quad \textit{and} \tag{13}$$

$$\operatorname{var}_\delta\Big(\frac{y}{x}\Big) = \frac{\operatorname{var}(y-\rho x)}{x_0^2}. \tag{14}$$

Proposition 4. *Let x_t, x_c, y_t, y_c be jointly distributed random variables with finite variances and means $x_{j,0} \neq 0$ and $y_{j,0}$ respectively, for $j \in \{t, c\}$. Let $\rho_j = y_{j,0}/x_{j,0}$. Then*

$$\operatorname{var}_\delta\Big(\frac{y_t}{x_t} \pm \frac{y_c}{x_c}\Big) = \frac{\operatorname{var}(y_t-\rho_t x_t)}{x_{t,0}^2} + \frac{\operatorname{var}(y_c-\rho_c x_c)}{x_{c,0}^2} \pm 2\frac{\operatorname{cov}(y_t-\rho_t x_t, y_c-\rho_c x_c)}{x_{t,0}x_{c,0}}.$$

Theorem 1. *Let $\mathcal{S}$ be an index set of finite cardinality $n > 0$. For $i \in \mathcal{S}$, let $W_{it} \sim \mathrm{Bern}(p_i)$ be independent with $0 < p_i < 1$ and set $W_{ic} = 1 - W_{it}$. Let*

$$\hat\tau = \frac{\sum_{i\in\mathcal{S}} W_{it}Y_{it}}{\sum_{i\in\mathcal{S}} W_{it}} - \frac{\sum_{i\in\mathcal{S}} W_{ic}Y_{ic}}{\sum_{i\in\mathcal{S}} W_{ic}}$$

where $(Y_{it}, Y_{ic}) \in [-B, B]^2$, for $B < \infty$. Then

$$\operatorname{var}_\delta(\hat\tau) = \frac{1}{n}\left(\frac{S_{tt}}{p_t^2} + \frac{S_{cc}}{p_c^2} + 2\frac{S_{tc}}{p_tp_c}\right) \tag{15}$$

where μ_t, μ_c, p_t, p_c, s_t, s_c, S_{tt}, S_{cc}, S_{tc} are defined at equations (10) *through* (12)*. If all $p_i \in [\epsilon, 1-\epsilon]$ for some $\epsilon > 0$, then*

$$\mathbb{E}_\delta(\hat\tau) = (\mu_t - \mu_c) + \Big(\frac{s_t}{p_t} - \frac{s_c}{p_c}\Big) + O\Big(\frac{1}{n}\Big). \tag{16}$$

Proof. See Section 6.1. □

The implied constant in $O(1/n)$ for equation (16) holds for all $n \geqslant 1$.

3.3 Delta method means and variances

We define the delta method bias of an estimate $\hat\tau_k$ via $\mathrm{Bias}_\delta(\hat\tau_k) = \mathbb{E}_\delta(\hat\tau_k) - \tau_k$.

Corollary 1. *Let $\hat\tau_{ok}$ be the ODB-only estimator from* (2)*. Then*

$$\operatorname{var}_\delta(\hat\tau_{ok}) = \frac{1}{n_{ok}}\left(\frac{S_{tt}}{p_t^2} + \frac{S_{cc}}{p_c^2} + 2\frac{S_{tc}}{p_tp_c}\right),$$

where s_t, s_c, p_t, p_c, S_{tt}, S_{cc} and S_{cc} are given in equations (9) *through* (12) *with $\mathcal{S} = \mathcal{O}_k$. If $1 < k < K$, then*

$$\mathrm{Bias}_\delta(\hat\tau_{ok}) = \frac{s_t}{p_t} - \frac{s_c}{p_c} + O\Big(\frac{1}{n_{ok}}\Big).$$

If also Assumption 4 holds, then

$$\mathrm{Bias}_\delta\,(\hat\tau_{ok}) = \frac{s_t}{p_t(1-p_t)} + O\biggl(\frac{1}{n_{ok}}\biggr).$$

Proof. For $1 < k < K$ we can apply Theorem 1 with $\epsilon = 1/K$. Under Assumption 4, $s_c = -s_t$, so the lead term in $\mathbb{E}_\delta(\hat\tau_k)$ is $s_t(1/p_t + 1/p_c) = s_t(p_t+p_c)/p_t(1-p_t) = s_t/p_t(1-p_t)$. □

Corollary 2. *Let $\hat\tau_{rk}$ be the RCT-only estimator from* (3)*. Then*

$$\mathrm{Bias}_\delta(\hat\tau_{rk}) = O\biggl(\frac{1}{n_{rk}}\biggr),$$

and

$$\begin{aligned}\mathrm{var}_\delta(\hat\tau_{rk}) &= \frac{\bar\sigma^2_{rk}}{n_{rk}p_r(1-p_r)}, \quad \textit{where}\\ \bar\sigma^2_{rk} &= \frac{1}{n_{rk}}\sum_{i\in\mathcal{R}_k}[(Y_{it}-\mu_{rkt})(1-p_r)+(Y_{ic}-\mu_{rkc})p_r]^2.\end{aligned} \tag{17}$$

where $\mu_{rkt} = \mu_t(\mathcal{R}_k)$ and $\mu_{rkc} = \mu_c(\mathcal{R}_k)$. Under Assumption 4, $\bar\sigma^2_{rk} = \sigma^2_{rkt} \equiv (1/n_{rk})\sum_{i\in\mathcal{R}_k}(Y_{it}-\mu_{rkt})^2$. If $p_r = 1/2$, then

$$\mathrm{var}_\delta(\hat\tau_{rk}) = \frac{1}{4n_k^2}\sum_{i\in\mathcal{R}_k}\biggl(\bar Y_i - \frac{\mu_{rkt}+\mu_{rkc}}{2}\biggr)^2$$

for $\bar Y_i = (Y_{it}+Y_{ic})/2$.

Proof. See Section 6.2. □

The RCT has a very tiny delta method bias which arises purely from the ratio estimator (random denominator) form of $\hat\tau_{rk}$. Conditional on there being at least one treated and one control subject in the stratum, it can be shown that $\hat\tau_{rk}$ is exactly unbiased rather than asymptotically unbiased. This follows from symmetry: at every value of $n_{rkt} \in \{1,2,\dots,n_{rk}-1\}$, the estimator is drawn uniformly at random from all permutations of the labels of who is treated and who is not, and unbiasedness follows.

Corollary 3. *Let $\hat\tau_{wk}$ be the weighted-average estimator* (5)*. Then, with $\lambda_k = n_{ok}/(n_{ok}+n_{rk})$,*

$$\mathrm{var}_\delta(\hat\tau_{wk}) = \frac{\lambda_k}{n_{ok}+n_{rk}}\biggl(\frac{S_{tt}}{p_t^2}+\frac{S_{cc}}{p_c^2}+2\frac{S_{tc}}{p_tp_c}\biggr) + \frac{1-\lambda_k}{n_{ok}+n_{rk}}\frac{\bar\sigma^2_{rk}}{p_r(1-p_r)},$$

where s_t, s_c, p_t, p_c, S_{tt}, S_{cc} and S_{cc} are given in equations (9) *through* (12) *with $\mathcal{S} = \mathcal{O}_k$, and $\bar\sigma^2_{rk}$ is defined at* (17)*. If $1 < k < K$, then*

$$\mathrm{Bias}_\delta(\hat\tau_{wk}) = \lambda_k\biggl(\frac{s_{okt}}{p_{okt}} - \frac{s_{okc}}{p_{okc}}\biggr) + O\biggl(\frac{1}{n_{ok}+n_{rk}}\biggr).$$

If Assumption 4 also holds, then

$$\mathrm{Bias}_\delta(\hat\tau_{wk}) = \frac{\lambda_k s_{okt}}{p_{okt}(1-p_{okt})} + O\left(\frac{1}{n_{ok}+n_{rk}}\right).$$

Proof. Using (7) and Corollaries 1 and 2, $\mathrm{Bias}_\delta(\hat\tau_{wk}) = \lambda_k \times \mathrm{Bias}_\delta(\hat\tau_{ok})$ for λ_k given in (5). This yields the lead terms in both expressions for $\mathrm{Bias}_\delta(\hat\tau_{wk})$. The error terms are $\lambda_k O(1/n_{ok}) = O(1/(n_{ok}+n_{rk}))$. Using independence of the RCT and ODB, Corollaries 1 and 2, and definition (8)

$$\begin{aligned}\mathrm{var}_\delta(\hat\tau_{wk}) &= \frac{\lambda_k^2}{n_{ok}}\left(\frac{S_{tt}}{p_t^2}+\frac{S_{cc}}{p_c^2}+2\frac{S_{tc}}{p_tp_c}\right)+(1-\lambda_k)^2\frac{\bar\sigma_{rk}^2}{n_{rk}p_r(1-p_r)}\\ &= \frac{\lambda_k}{n_{ok}+n_{rk}}\left(\frac{S_{tt}}{p_t^2}+\frac{S_{cc}}{p_c^2}+2\frac{S_{tc}}{p_tp_c}\right)+\frac{1-\lambda_k}{n_{ok}+n_{rk}}\frac{\bar\sigma_{rk}^2}{p_r(1-p_r)}\end{aligned}\qquad\square$$

We see that the weighted method reduces the delta method variance of the ODB-only estimate by a factor of λ_k. This will not typically be a large reduction when the the ODB is much larger than the RCT.

The spiked-in estimator's bias and variance cannot be computed as a corollary of Theorem 1, but they can be computed directly.

Corollary 4. *Let $\hat\tau_{sk}$ be the spiked-in estimator* (5)*. Then*

$$\mathrm{var}_\delta(\hat\tau_{sk}) = \frac{1}{n_{ok}+n_{rk}}\left(\frac{S_{tt}}{p_t^2}+\frac{S_{cc}}{p_c^2}+2\frac{S_{tc}}{p_tp_c}\right),$$

where s_t, s_c, p_t, p_c, S_{tt}, S_{cc} and S_{tc} are given in equations (9) *through* (12) *with $\mathcal{S} = \mathcal{O}_k \cup \mathcal{R}_k$. If $1 < k < K$, then*

$$\mathrm{Bias}_\delta\left(\hat\tau_{sk}\right) = \frac{s_t}{p_t} - \frac{s_c}{p_c} + O\left(\frac{1}{n_{ok}+n_{rk}}\right).$$

If Assumption 4 also holds, then

$$\mathrm{Bias}_\delta\left(\hat\tau_{sk}\right) = \frac{s_t}{p_t(1-p_t)} + O\left(\frac{1}{n_{ok}+n_{rk}}\right).$$

Proof. The spike-in estimates are computed by pooling $\mathcal{O}_k$ and $\mathcal{R}_k$ into their union. □

The edge bins are not covered by Corollary 4. Inspection of the proof of Theorem 1 shows that the bias error term is $O((n_{ok}+n_{rk})/n_{rk}^2)$. No such bound is available for $\hat\tau_{ok}$ or $\hat\tau_{wk}$ for edge bins.

To relate the bias of $\hat\tau_{sk}$ to that of the other estimators, we write it in terms of the quantities computed using $\mathcal{S} = \mathcal{O}_k$ and $\mathcal{S} = \mathcal{R}_k$. Denoting these quantities using an additional subscript of o and r,

$$\begin{aligned}\mathrm{Bias}_\delta\left(\hat\tau_{sk}\right) = \Delta_k n_{ok}&\left(\frac{p_{okt}}{n_{ok}p_{okt}+n_{rk}p_{rkt}} - \frac{p_{okc}}{n_{ok}p_{okc}+n_{rk}p_{rkc}}\right) +\\ &s_{okt}\frac{n_{ok}}{n_{ok}p_{okt}+n_{rk}p_{rkt}} - s_{okc}\frac{n_{ok}}{n_{ok}p_{okc}+n_{rk}p_{rkc}} + O\left(\frac{1}{n_{ok}+n_{rk}}\right).\end{aligned}\tag{18}$$

The bias for $\hat{\tau}_{rk}$ is zero. The bias for $\hat{\tau}_{ok}$ has terms analogous to the second and third (and error) terms above, but the first term is new to $\hat{\tau}_{sk}$. This term is linear in Δ_k. For large values of Δ_k, this term will dominate, yielding biases that can easily exceed those of $\hat{\tau}_{ok}$. This is the fundamental danger of the spiked-in estimator: if the mean potential outcomes differ substantially between ODB and RCT subjects with similar value of the propensity score function, then the estimation will be poor due to large bias.

3.4 The dynamic weighted estimator

The bias-variance tradeoffs are intrinsically different in each stratum. Using results from the prior section, we derive a dynamic weighted estimator that uses different weights in each stratum. Our dynamic weighted estimator is based on Assumption 4, though we will test it in settings where that assumption does not hold.

From Proposition 1, the MSE-optimal convex combination of $\hat{\tau}_{ok}$ and $\hat{\tau}_{rk}$ is $c_{*k}\hat{\tau}_{ok} + (1 - c_{*k})\hat{\tau}_{rk}$ where $c_{*k} = \mathrm{var}(\hat{\tau}_{rk})/(\mathrm{var}(\hat{\tau}_{rk}) + \mathrm{MSE}(\hat{\tau}_{ok}))$. The dynamic weighted estimator is

$$\hat{\tau}_{dk} = \hat{c}_{*k}\hat{\tau}_{ok} + (1 - \hat{c}_{*k})\hat{\tau}_{rk}, \quad \text{with} \quad \hat{c}_{*k} = \frac{\widehat{\mathrm{var}}(\hat{\tau}_{rk})}{\widehat{\mathrm{var}}(\hat{\tau}_{rk}) + \widehat{\mathrm{MSE}}(\hat{\tau}_{ok})}, \tag{19}$$

for plug-in estimators of $\mathrm{MSE}(\hat{\tau}_{ok})$ and $\mathrm{var}(\hat{\tau}_{rk})$. To obtain our MSE estimates we use $\widetilde{\mathrm{MSE}}(\cdot) = \mathrm{Bias}_{\delta}(\cdot)^2 + \mathrm{var}_{\delta}(\cdot)$ taking the delta method moments from Corollaries 1 and 2. These expressions include some unknown population quantities that we then approximate from the data to get $\widehat{\mathrm{MSE}}(\cdot)$.

For the ODB estimate we use

$$\widetilde{\mathrm{MSE}}(\hat{\tau}_{ok}) = \left(\frac{s_t}{p_t(1 - p_t)}\right)^2 + \frac{1}{n_{ok}}\left(\frac{S_{tt}}{p_t^2} + \frac{S_{cc}}{p_c^2} + 2\frac{S_{tc}}{p_t p_c}\right)$$

where the quantities on the right hand side are given in Section 3.1 with $\mathcal{S} = \mathcal{O}_k$. For the RCT estimate we use

$$\widetilde{\mathrm{var}}(\hat{\tau}_{rk}) = \frac{\bar{\sigma}^2_{rk}}{p_r(1 - p_r)n_{rk}}, \quad \text{with} \quad \bar{\sigma}^2_{rk} = \frac{1}{n_{rk}}\sum_{i \in \mathcal{R}_k} W_{it}\hat{\sigma}^2_{rkt} + W_{ic}\hat{\sigma}^2_{rkc}$$

where $\hat{\sigma}^2_{rkt}, \hat{\sigma}^2_{rkc}$ are the sample variances observed among the treated and control units respectively. Both of these estimates use Assumption 4.

The values of p_t and p_c are known: $p_t = \sum_{i \in \mathcal{O}_k} p_{it}/n_{ok}$ where p_{it} is the propensity $e(\boldsymbol{x}_i)$ and $p_c = 1 - p_t$. We use Horvitz-Thompson style inverse probability weighting to estimate other quantities, as follows:

$$\hat{\rho}_t = \frac{\sum_{i \in \mathcal{O}_k} W_{it}Y_{it}}{\sum_{i \in \mathcal{O}_k} W_{it}}, \qquad \hat{\rho}_c = \frac{\sum_{i \in \mathcal{O}_k} W_{ic}Y_{ic}}{\sum_{i \in \mathcal{O}_k} W_{ic}},$$

$$
\begin{aligned}
\hat{s}_t &= \frac{\sum_{i\in\mathcal{O}_k} W_{it}}{n_{ok}} \left(\sum_{i\in\mathcal{O}_k} W_{it}Y_{it} - p_t \sum_{i\in\mathcal{O}_k} W_{it}Y_{it}/p_{it} \right) \\
&\qquad + \frac{\sum_{i\in\mathcal{O}_k} W_{ic}}{n_{ok}} \left(\sum_{i\in\mathcal{O}_k} W_{ic}Y_{ic} - p_c \sum_{i\in\mathcal{O}_k} W_{ic}Y_{ic}/p_{ic} \right), \\
\hat{S}_{tt} &= \frac{\sum_{i\in\mathcal{O}_k} W_{it}p_{it}(1-p_{it})(Y_{it}-\hat{\rho}_t)^2}{\sum_{i\in\mathcal{O}_k} W_{it}}, \quad \text{and} \\
\hat{S}_{cc} &= \frac{\sum_{i\in\mathcal{O}_k} W_{ic}p_{it}(1-p_{it})(Y_{ic}-\hat{\rho}_c)^2}{\sum_{i\in\mathcal{O}_k} W_{ic}}.
\end{aligned}
$$

The sole quantity that does not have a Horvitz-Thompson estimator is $S_{tc}(\mathcal{O}_k)$, because we never observe both the potential outcomes for a given unit. First, we write S_{tc} as

$$
\frac{1}{n}\sum_{i\in\mathcal{O}_k} W_{it}p_{it}(1-p_{it})(Y_{it}-\rho_t)(Y_{ic}-\rho_c) + \frac{1}{n}\sum_{i\in\mathcal{O}_k} W_{ic}p_{it}(1-p_{it})(Y_{it}-\rho_t)(Y_{ic}-\rho_c).
$$

Next, under Assumption 4,

$$
Y_{it}-\rho_t = Y_{ic}+\tau_k-\mu_t-s_t/p_t = Y_{ic}-\rho_c-\frac{s_t}{p_tp_c},
$$

and similarly $Y_{ic}-\rho_c = Y_{it}-\rho_t+s_t/(p_tp_c)$. Therefore

$$
\begin{aligned}
S_{tc} = \frac{1}{n}\sum_{i\in\mathcal{O}_k} W_{it}p_{it}(1-p_{it})(Y_{it}-\rho_t)^2 + \frac{1}{n}\sum_{i\in\mathcal{O}_k} W_{ic}p_{it}(1-p_{it})(Y_{ic}-\rho_c)^2 - \\
- \frac{s_t}{np_t(1-p_t)}\left(\sum_{i\in\mathcal{O}_k} W_{it}p_{it}(1-p_{it})(Y_{it}-\rho_t) - W_{ic}p_{it}(1-p_{it})(Y_{ic}-\rho_c)\right)
\end{aligned}
\tag{20}
$$

and we get $\hat{S}_{tc}$ by plugging the above estimates of ρ_t, ρ_c and known values of p_t, p_c into (20). Although Assumption 4 is used to derive the estimator, some of our simulations will test it under a violation of that assumption.

3.5 Performance comparison

The ideal dynamic estimator with the optimal weight c_{k*} must be at least as good as $\hat{\tau}_{ok}$, $\hat{\tau}_{rk}$ and $\hat{\tau}_{wk}$ because those estimators are all special cases of weighting estimators belonging to the class that c_{k*} optimizes over. Our estimator $\hat{\tau}_{dk}$ will not always be better than those other estimators, because it uses an estimate $\hat{c}_{k*}$ which could introduce enough error to make it less efficient.

When combining stratum-based estimates $\hat{\tau}_k$ into the weighted estimator $\hat{\tau} = \sum_k \omega_k\hat{\tau}_k$, there is the possibility of biases canceling between strata. None of the competing estimators we consider are designed to exploit such cancellation. For large strata, c_{k*} should be well estimated. To arrange cancellations among biased within-stratum estimates would require domain-specific assumptions that we do not make here.

The comparison to the spiked-in estimator is more complex. As we saw in equation (18), the bias can grow without bound in Δ_k, so for large Δ_k this estimator will have the largest MSE. However, for small values of Δ_k, the spiked-in estimator can outperform all the other estimators. To see why, we make a direct comparison with the dynamic weighted estimator and reference our prior discussion showing the dynamic weighted estimator will generally outperform $\hat{\tau}_{ok}, \hat{\tau}_{rk}$ and $\hat{\tau}_{wk}$.

After some algebra, the difference between spiked-in and RCT estimators $\hat{\tau}_{sk} - \hat{\tau}_{rk}$ is

$$c_{kt}\left(\frac{\sum_{i\in\mathcal{O}_k} W_{it}Y_{it}}{\sum_{i\in\mathcal{O}_k} W_{it}} - \frac{\sum_{i\in\mathcal{R}_k} W_{it}Y_{it}}{\sum_{i\in\mathcal{R}_k} W_{it}}\right) - c_{kc}\left(\frac{\sum_{i\in\mathcal{O}_k} W_{ic}Y_{ic}}{\sum_{i\in\mathcal{O}_k} W_{ic}} - \frac{\sum_{i\in\mathcal{R}_k} W_{ic}Y_{ic}}{\sum_{i\in\mathcal{R}_k} W_{ic}}\right).$$

where

$$c_{kt} = \frac{\sum_{i\in\mathcal{O}_k} W_{it}}{\sum_{i\in\mathcal{O}_k} W_{it} + \sum_{i\in\mathcal{R}_k} W_{it}} \quad \text{and} \quad c_{kc} = \frac{\sum_{i\in\mathcal{O}_k} W_{ic}}{\sum_{i\in\mathcal{O}_k} W_{ic} + \sum_{i\in\mathcal{R}_k} W_{ic}}$$

are the empirical ratios of treated and control point counts from the ODB and RCT. By comparison, $\hat{\tau}_{dk} - \hat{\tau}_{rk}$ equals

$$c_{k\star}\left(\frac{\sum_{i\in\mathcal{O}_k} W_{it}Y_{it}}{\sum_{i\in\mathcal{O}_k} W_{it}} - \frac{\sum_{i\in\mathcal{R}_k} W_{it}Y_{it}}{\sum_{i\in\mathcal{R}_k} W_{it}}\right) - c_{k\star}\left(\frac{\sum_{i\in\mathcal{O}_k} W_{ic}Y_{ic}}{\sum_{i\in\mathcal{O}_k} W_{ic}} - \frac{\sum_{i\in\mathcal{R}_k} W_{ic}Y_{ic}}{\sum_{i\in\mathcal{R}_k} W_{ic}}\right).$$

An oracle would choose $c_{k\star}$ using Proposition 1. The dynamic estimator uses a plug-in principle to estimate the oracle's $c_{k\star}$. Here we see that the oracle is working in a one parameter family for each bin k, while the spiked-in estimator has a pair of weights c_{kt} and c_{kc} that are not necessarily within the family that the oracle optimizes over.

4 Simulations

Our goal is to estimate the average treatment effect in the target population, from which we assume the ODB data was randomly sampled. The value of the RCT is that it can substitute for ODB data in places where that data is sparse due to the treatment assignment mechanism.

We simulate two high level scenarios. In one, the RCT is a random sample from the same population that the ODB came from. Then the RCT and ODB data differ only in their treatment assignment mechanisms. We consider this case the ideal one for our approach of merging the RCT into the ODB. In the other scenario, the RCT is subject to some potentially biasing inclusion criteria on the explanatory variables $\boldsymbol{x}_i$. Such biases are a frequent concern for RCTs (Susukida et al., 2016; Stuart and Rhodes, 2017).

γ:	$\begin{pmatrix}1\\1\\1\\0\\0\end{pmatrix}$	$\begin{pmatrix}\sqrt{2}\\\sqrt{2}\\\sqrt{2}\\0\\0\end{pmatrix}$	$\begin{pmatrix}\sqrt{3}/2\\-\sqrt{3}/2\\\sqrt{3}/2\\-\sqrt{3}/2\\0\end{pmatrix}$	$\begin{pmatrix}\sqrt{6}/2\\-\sqrt{6}/2\\\sqrt{6}/2\\-\sqrt{6}/2\\0\end{pmatrix}$
$\gamma^{\mathsf{T}}\beta$:	3	$3\sqrt{2}$	0	0
$\Vert\gamma\Vert^2$:	3	6	3	6

Table 2: These are the four γ vectors used in our simulations. The first two correlate with the mean response vector β, while the second two do not. The second and fourth imply larger sampling biases than the first and third do.

For both of these high level scenarios, we vary the treatment effect over strata, making it either constant, linear or quadratic in $k = 1, \dots, K$. Section 4.1 shows results for our ideal case where $\boldsymbol{x}_i$ have the same distribution in both data sets and Assumption 4 holds. Section 4.2 models a sampling bias for the $\boldsymbol{x}_i$ values in the RCT while retaining Assumption 4. Section 4.3 removes Assumption 4 from both of the prior simulation settings.

4.1 Simulation of the ideal case

We begin with the simulations satisfying Assumption 4, with the RCT sampled from the same distribution as the ODB. This is an ideal case. First we describe how the ODB data are generated, then the RCT data.

In all of our simulations $\boldsymbol{x}_i \in \mathbb{R}^5$. The ODB has $n_o = 5{,}000$ subjects. We generate $\boldsymbol{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_5)$ for $i \in \mathcal{O}$ and we assume that for the control condition

$$Y_{ic} = \boldsymbol{x}_i^{\mathsf{T}}\beta + \varepsilon_i, \quad \text{for } \beta = (1,1,1,1,1)^{\mathsf{T}}$$

where ε_i are generated as IID $\mathcal{N}(0,1)$ random variables. We assume that the user does not know the precise form of the generative model and uses the stratified estimates we presented above.

The treatment variables in the ODB are independent Bernoulli random variables with

$$\Pr(W_i = 1) = \frac{1}{1 + e^{-\gamma^{\mathsf{T}}\boldsymbol{x}_i}}.$$

We consider the four γ vectors given in Table 2. Two of them are orthogonal to β. The others are correlated with β and will result in the test group having higher average responses in the ODB than the control group. For each correlation pattern we have two sizes of $\Vert\gamma\Vert$.

The treatment values Y_{it} are equal to Y_{ic} plus a treatment effect that obeys Assumption 4. We use three structures. In all cases $Y_{it} = Y_{ic} + \tau_k$ for $i \in \mathcal{O}_k$. The values of τ_k in constant, linear and quadratic treatment effect models are

$$\tau_k = T, \quad \tau_k = T \times \frac{k}{K}, \quad \text{and} \quad \tau_k = T \times \Big(\frac{k}{K} - \frac{1}{2}\Big)^2 \tag{21}$$

respectively. In each case we choose the scale $T > 0$ so that Cohen's d (Cohen, 1988) in the ODB precisely equals 0.5, which Cohen calls a medium effect size. The value of T varies across simulations based on the simulation settings and the randomness in the sampling of $\boldsymbol{x}_i$.

We turn now to the RCT. It has $n_r = 200$ subjects in it. Their $\boldsymbol{x}_i$ are chosen from the same distribution as in the ODB but now $W_i \overset{\text{iid}}{\sim} \text{Bern}(1/2)$. The same constant, linear and quadratic treatment effects from the ODB are used in the RCT.

We simulate the covariates $\boldsymbol{x}_i$ and potential outcomes (Y_{it}, Y_{ic}) for $i \in \mathcal{O} \cup \mathcal{R}$ 100 times. For each realization of the covariates and potential outcomes we make 20 independent simulations of the treatment variables W_i, for a total of 2,000 simulations. In all cases we chose $K = 20$ propensity bins where the k'th one has $e(\boldsymbol{x}_i) \in [(k-1)/20, k/20)$ for $k = 1, \dots, K$. This simulation satisfies Assumptions 1 through 4.

In each simulation run, we estimate the average treatment effect using each of our five estimators. We also estimate using an 'oracle' estimator, which knows the true MSE of the ODB-only and RCT-only estimators in each stratum and thus the optimal weighting between these estimators. The MSE of the estimators are computed across all 2,000 simulation runs,.

The results are shown in Table 3. In this ideal setting, the spike-in estimator always has the lowest MSE. It was always superior to the better of the RCT and ODB estimators. Even though the RCT only adds 200 subjects to the 5,000 in the ODB, it leads to a spiked-in estimator whose MSE ranges from about 45% to about 80% of that of the ODB. As we discussed in Section 3.5, this setting has $\Delta_k = 0$ and the RCT and ODB points have similar variances conditionally on the propensity score. It even outperforms the oracle estimator which optimizes the relative weights on the RCT and ODB within strata. The spiked-in estimator can beat the oracle because it is not one of the weighting schemes over which the oracle has optimized. The oracle is the second-best performer in each condition. The dynamic weighted estimator – which seeks to recover the oracle weights – has an MSE only slightly inflated relative to the oracle, with no performance gap larger than 15%. The dynamic weighted estimator also generally outperforms the weighted-average estimator.

Trt.	Cor.	$\|\gamma\|_2^2$	ODB	RCT	Wtd.	Spike	Dyn.	Oracle
c	y	3	0.0077	0.0742	0.0072	0.0053	0.0067	0.0059
c	y	6	0.0230	0.0808	0.0212	0.0103	0.0138	0.0134
c	n	3	0.0122	0.1443	0.0116	0.0094	0.0120	0.0104
c	n	6	0.0209	0.1538	0.0196	0.0138	0.0172	0.0158
l	y	3	0.0076	0.0750	0.0072	0.0053	0.0067	0.0060
l	y	6	0.0225	0.0784	0.0209	0.0112	0.0139	0.0134
l	n	3	0.0123	0.1377	0.0116	0.0098	0.0117	0.0104
l	n	6	0.0220	0.1524	0.0204	0.0137	0.0175	0.0162
q	y	3	0.0073	0.0799	0.0069	0.0052	0.0066	0.0058
q	y	6	0.0217	0.0751	0.0201	0.0101	0.0138	0.0132
q	n	3	0.0127	0.1496	0.0120	0.0093	0.0119	0.0107
q	n	6	0.0214	0.1503	0.0201	0.0127	0.0176	0.0160

Table 3: MSEs for treatment effect in the ideal setting. Column 1 gives treatment (constant, linear, quadratic). Column 2 shows whether the propensity was correlated with the mean response. Column 3 indicates the magnitude of the propensity vector γ. The remaining columns are mean squared errors for the overall treatment from our 5 estimators and an oracle. In every case, the spiked-in estimator using (4) has lowest MSE.

Estimator Performance: Quadratic Treatment Effect, Ideal Case

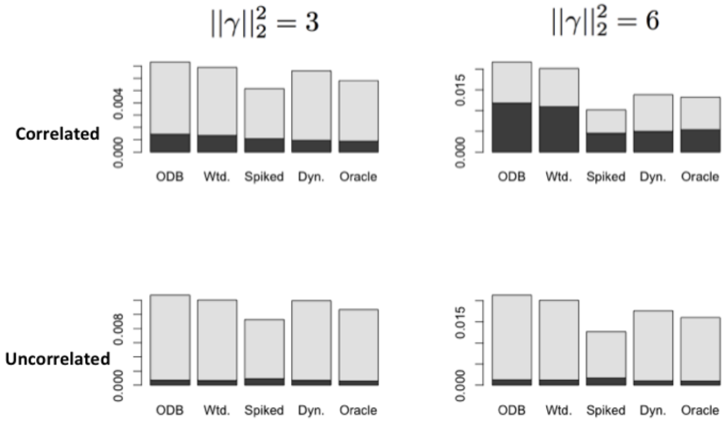

Figure 1: Performance measures across all 2,000 simulations run in the ideal case. Bias squared is shown in black, and variance in gray, so that total bar height represents the MSE. The much larger values for the RCT estimator are excluded to make visual comparison easier.

Trt.	**Cor.**	$\|\gamma\|_2^2$	**ODB**	**RCT**	**Wtd.**	**Spike**	**Dyn.**	**Oracle**
c	y	3	0.0078	0.1405	0.0072	0.0129	0.0064	0.0061
c	y	6	0.0218	0.1759	0.0198	0.0182	0.0131	0.0123
c	n	3	0.0114	0.1652	0.0104	0.0330	0.0100	0.0091
c	n	6	0.0211	0.1873	0.0196	0.0648	0.0175	0.0152
l	y	3	0.0076	0.3495	0.0069	0.0120	0.0060	0.0056
l	y	6	0.0220	0.4347	0.0200	0.0176	0.0130	0.0123
l	n	3	0.0120	0.2392	0.0111	0.0344	0.0105	0.0095
l	n	6	0.0216	0.2782	0.0197	0.0645	0.0169	0.0147
q	y	3	0.0076	0.3641	0.0069	0.0123	0.0062	0.0058
q	y	6	0.0214	0.2883	0.0193	0.0190	0.0127	0.0120
q	n	3	0.0120	0.3042	0.0112	0.0330	0.0111	0.0099
q	n	6	0.0209	0.2371	0.0193	0.0675	0.0170	0.0151

Table 4: MSEs for treatment effect in the setting with restricted enrollments. The columns are the same as in Table 3. Here the oracle estimator is always best and the dynamic estimator is the best of the ones that can be implemented.

The outcomes in Table 3 are quite consistent. Ten out of 12 settings have the same ordering. From best to worst they are: spiked, oracle, dynamic, weighted, ODB and RCT. In two of the cases (rows 3 and 7) the weighted method very slightly outperforms the dynamic method.

Inspection of the data in Table 3 shows that the RCT alone is far from competitive. This is not surprising as it has only a small amount of data. A graphical investigation is presented in Figure 1. Bias squared is shown in black, and variance in gray, so that total bar height represents the MSE. The treatment patterns make very little difference, so we show only the case of the quadratic treatment effect. To make comparison easier, we exclude the RCT estimator. These results show that the advantage of the spiked-in estimator is greatest when $\|\gamma\|$ is large. In this case, the benefit mostly accrues due to a reduced variance for the spiked-in estimator relative to the dynamic estimator.

4.2 Restrictive enrollment criteria

It is common for an RCT to have enrollment criteria such that the values of $\boldsymbol{x}_i$ in it are different from those in the general population. The RCT might be designed to avoid frail patients. Or it might be designed to include patients with the worst prognoses, who are most in need of a better treatment. We illustrate restrictive enrollment by having the RCT sample $\boldsymbol{x}_i$ from $\mathcal{N}(0, I)$ subject to both $x_{i1} < -1$ and $x_{i5} < -1$. Because our β vector has all positive entries, these restrictions mean that subjects in the RCT tend to have smaller values of Y_{ic} than those in the ODB. Smaller could either mean better or worse depending on what quantity Y measures.

Estimator Performance: Quadratic Treatment Effect, Restricted Enrollment Case

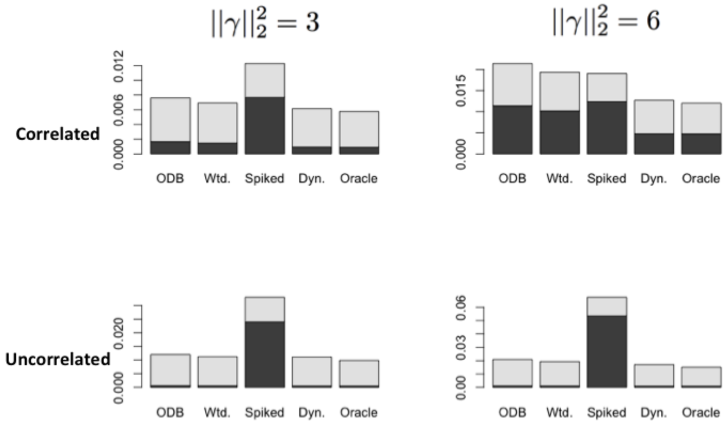

Figure 2: Performance measures across all 2,000 simulations run in the restricted enrollment case. Bias squared is shown in black, and variance in gray, so that total bar height represents the MSE. The much larger values for the RCT estimator are excluded to make visual comparison easier.

The results under this restriction are presented in Table 4. Nine of the 12 settings have the same ordering. From best to worst they are: oracle, dynamic, weighted, ODB, spiked, and RCT. In the remaining three cases, the spiked estimator is third, the weighted estimator fourth, and the ODB estimator fifth, with the other rankings unchanged. Across all settings, the best usable method is the dynamic one. The dynamic MSE was never more than 15% higher than that of the oracle that it seeks to approximate. Sometimes it was up to 40% more efficient than the ODB-only estimator.

The bad performance of the spiked-in estimator here is attributable to the restriction on the fifth component of $\boldsymbol{x}_i$. That restriction affects the outcomes (Y_{ic}, Y_{it}) but not the propensity score because $\gamma_5 = 0$. The result is a large Δ_k in each stratum, making the spiked-in estimator perform much worse than in the prior scenario. This effect can be seen in Figure 2, where the spiked-in estimator consistently demonstrates a large squared bias, resulting in a high MSE.

4.3 Violation of Assumption 4

In this section we simulate in a setting where Assumption 4 fails to hold but Assumption 3 does hold. We modify the linear and quadratic treatment effects in equation (21) to have

$$Y_{it} - Y_{ic} = T \times e(\boldsymbol{x}_i), \quad \text{and} \quad Y_{it} - Y_{ic} = T \times (e(\boldsymbol{x}_i) - 1/2)^2$$

Trt.	Cor.	$\|\gamma\|_2^2$	ODB	RCT	Wtd.	Spike	Dyn.	Oracle
l	y	3	0.0080	0.0751	0.0075	0.0056	0.0070	0.0063
l	y	6	0.0232	0.0778	0.0215	0.0108	0.0140	0.0136
l	n	3	0.0121	0.1545	0.0115	0.0090	0.0118	0.0103
l	n	6	0.0216	0.1430	0.0201	0.0132	0.0173	0.0156
q	y	3	0.0072	0.0767	0.0067	0.0051	0.0066	0.0059
q	y	6	0.0193	0.0780	0.0180	0.0100	0.0127	0.0122
q	n	3	0.0128	0.1582	0.0120	0.0094	0.0118	0.0107
q	n	6	0.0214	0.1455	0.0199	0.0135	0.0169	0.0155

Table 5: These are the results of the simulations where Assumption 4 is violated but the RCT has the same $\boldsymbol{x}$ distribution as the ODB.

respectively. T is again selected so that Cohen's d in the ODB is equal to 0.5. The treatment difference now depends on the actual propensity of each subject but it varies with the strata. We do not re-simulate the constant case because it is the same either way.

If the RCT is sampled randomly from the population, we get the results in Table 5. Here, again, the rankings are very stable. Seven times out of 8, the ranking from best to worst is: spiked, oracle, dynamic, weighted, ODB and RCT. One time the weighted estimator slightly outperformed the dynamic estimator. The orderings are essentially unchanged from the case when Assumption 4 held. The value of Δ_k here, while not zero, is not very large. The dynamic estimator is still a good approximation to the oracle, with an MSE never more than 14% larger.

Finally, we consider the setting where Assumption 4 is violated and the RCT has the enrollment restrictions from Section 4.2. The results are in Table 6. In 7 of 8 cases, the ranking is: oracle, dynamic, weighted, ODB, spiked and RCT, just as it predominantly was when Assumption 4 held. The single dissimilar case still has the oracle and dynamic estimators as the top performers. Taken together, these results indicate that the dynamic weighted estimator is robust to this type of weakening of Assumption 4.

5 Conclusions

We have developed some propensity based methods to merge data from a randomized controlled trial with data on the same phenomenon from an observational data base. Our goal is to reduce the mean squared error of the overall population treatment effect.

The strategies we use are based on the propensity that the RCT data would have had, had they been in the ODB. The simplest strategy is to spike the RCT data into the corresponding propensity strata. It works well in theory and in experiments when the covariate distribution in the RCT matches that of the

Trt.	**Cor.**	$\|\gamma\|_2^2$	**ODB**	**RCT**	**Wtd.**	**Spike**	**Dyn.**	**Oracle**
l	y	3	0.0078	0.3984	0.0071	0.0122	0.0064	0.0059
l	y	6	0.0237	0.4697	0.0214	0.0183	0.0138	0.0131
l	n	3	0.0117	0.2776	0.0108	0.0338	0.0103	0.0094
l	n	6	0.0207	0.2568	0.0189	0.0647	0.0160	0.0145
q	y	3	0.0067	0.2792	0.0062	0.0136	0.0058	0.0054
q	y	6	0.0187	0.2359	0.0171	0.0202	0.0123	0.0116
q	n	3	0.0116	0.2518	0.0106	0.0366	0.0101	0.0092
q	n	6	0.0207	0.2229	0.0189	0.0675	0.0162	0.0145

Table 6: These are the results of the simulations where Assumption 4 is violated and the $\boldsymbol{x}$ in the RCT are subject to restrictive enrollment critieria.

ODB. If however those distributions differ sharply, as they could for an RCT with restrictive enrollments, then the spiked-in estimator can perform very badly and even be worse than using the ODB alone without the RCT. We developed an alternative estimator based on taking a weighted average of the ODB and RCT data within every stratum. An oracle knowing the biases and variances of the ODB and RCT within each stratum could make a principled choice of weight vector. We developed an estimator that uses the plug-in principle to estimate that weight vector. On biased examples, it greatly outperformed both the spike-in estimator and the ODB itself. Our conclusion is that the spike-in estimator is the best choice when the RCT covariates come from the same distribution as the ODB, but otherwise we prefer the dynamic weighted estimator.

Acknowledgments

Evan Rosenman was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. This work was also supported by the NSF under grants DMS-1521145 and DMS-1407397.

References

Brand, J. E. and Davis, D. (2011). The impact of college education on fertility: Evidence for heterogeneous effects. *Demography*, 48(3):863–887.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition.

DellaPosta, D. J. (2013). The heterogeneous economic returns to military service: Evidence from the Wisconsin longitudinal study. *Research in Social Stratification and Mobility*, 34:73–95.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2013). From SATE to PATT: Combining experimental with observational studies to estimate population treatment effects.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA.

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, 150(4):327–333.

Peysakhovich, A. and Lada, A. (2016). Combining observational and experimental data to find heterogeneous treatment effects. *CoRR*, abs/1611.02385.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Stuart, E. A. and Rhodes, A. (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Evaluation review*, 41(4):357–388.

Stuart, E. A. and Rubin, D. B. (2007). Best practices in quasi-experimental designs: Matching methods for causal inference. In Osborne, J., editor, *Best Practices in Quantitative Social Science*, Thousand Oaks, CA. Sage Publications.

Susukida, R., Crum, R. M., Stuart, E. A., Ebnesajjad, C., and Mojtabai, R. (2016). Assessing sample representativeness in randomized controlled trials: application to the national institute of drug abuse clinical trials network. *Addiction*, 111(7):1226–1234.

Xie, Y., Brand, J. E., and Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347.

Xu, Z. and Kalbfleisch, J. D. (2010). Propensity score matching in randomized clinical trials. *Biometrics*, 66(3):813–823.

6 Appendix

This appendix contains two of the lengthier proofs.

6.1 Proof of Theorem 1

First let $x_t = \sum_{i\in\mathcal{S}} W_{it}$ and $y_t = \sum_{i\in\mathcal{S}} W_{it}Y_{it}$. From Proposition 3,

$$\mathbb{E}_\delta\Bigl(\frac{y_t}{x_t}\Bigr) = \frac{y_{t,0}}{x_{t,0}} - \frac{\mathrm{cov}(y_t - \rho_t x_t, x_t)}{x_{t,0}^2}$$

with $x_{t,0} = \mathbb{E}(x_t)$, $y_{t,0} = \mathbb{E}(y_t)$ and $\rho_t = y_{t,0}/x_{t,0}$. Next $x_{t,0} = np_t$ and $y_{t,0} = n(s_t + \mu_t p_t)$. Therefore $\rho_t = \mu_t + s_t/p_t$. By independence of the W_{it},

$$\begin{aligned}\mathrm{cov}(y - \rho x, x) &= \sum_{i\in\mathcal{S}} \mathrm{cov}(W_{it}(Y_{it} - \rho), W_{it}) = \sum_{i\in\mathcal{S}}(Y_{it} - \rho)p_i(1 - p_i)\\ &= -\sum_{i\in\mathcal{S}}(Y_{it} - \rho)p_i^2 = O(n),\end{aligned}$$

because $|Y_{it}| \leqslant B$. Furthermore, $x_{t,0}^2 \geqslant \epsilon^2 n^2$ and so

$$\mathbb{E}_\delta\Bigl(\frac{y_t}{x_t}\Bigr) = \mu_t + \frac{s_t}{p_t} + O\Bigl(\frac{1}{n}\Bigr).$$

Applying the same argument to the second term in $\hat\tau$ establishes equation (16) for $\mathbb{E}_\delta(\hat\tau)$.

Now we turn to the delta method variance, using Proposition 4. By independence of W_i,

$$\mathrm{var}(y_t - \rho_t x_t) = \sum_{i\in\mathcal{S}} \mathrm{var}(W_i(Y_{it} - \rho_t)) = \sum_{i\in\mathcal{S}} p_i(1 - p_i)(Y_{it} - \rho_t)^2 = nS_{tt}(\mathcal{S})$$

and by the same argument, $\mathrm{var}(y_c - \rho_c x_c) = nS_{cc}(\mathcal{S})$. Next

$$\mathrm{cov}(y_t - \rho_t x_t, y_c - \rho_c x_c) = \sum_{i\in\mathcal{S}} \mathrm{cov}(W_i(Y_{it} - \rho_t), (1 - W_i)(Y_{ic} - \rho_c)) = -nS_{tc}(\mathcal{S})$$

because $\mathrm{cov}(W_i, 1 - W_i) = -p_i(1 - p_i)$. The denominators in Proposition 4 simplify to $n^2p_t^2$, $n^2p_c^2$, and $n^2p_tp_c$. Then

$$\mathrm{var}_\delta\Bigl(\frac{y_t}{x_t} - \frac{y_c}{x_c}\Bigr) = \frac{1}{n}\Bigl(\frac{S_{tt}}{p_t^2} + \frac{S_{cc}}{p_c^2} - 2\frac{-S_{tc}}{p_tp_c}\Bigr)$$

completing the proof of (15).

6.2 Proof of Corollary 2

The RCT sampling probabilities are all p_r. Theorem 1 applies with $\epsilon = \min(p_r, 1-p_r)$. Because the sampling probability is the same for all subjects i, the covariances $s_t(\mathcal{R}_k)$ and $s_c(\mathcal{R})$ from equation (11) vanish, making $\mathbb{E}_\delta(\hat\tau_{rk}) = O(1/n_{rk})$. Next from Theorem 1,

$$\mathrm{var}_\delta(\hat\tau) = \frac{1}{n_{rk}}\left(\frac{S_{tt}}{p_t^2} + \frac{S_{cc}}{p_c^2} + 2\frac{S_{tc}}{p_tp_c}\right)$$

with parts defined using $\mathcal{S} = \mathcal{R}_k$. Then $p_t = p_r$, $p_c = 1 - p_r$, $\rho_t = \mu_t$, $\rho_c = \mu_c$, $S_{tt} = [p_r(1-p_r)/n_{rk}] \sum_{i\in\mathcal{R}_k}(Y_{it}-\mu_t)^2$, $S_{cc} = [p_r(1-p_r)/n_{rk}] \sum_{i\in\mathcal{R}_k}(Y_{ic}-\mu_c)^2$, and $S_{tc} = [p_r(1-p_r)/n_{rk}] \sum_{i\in\mathcal{R}_k}(Y_{it}-\mu_t)(Y_{ic}-\mu_c)$. Making these substitutions,

$$\begin{aligned}
\operatorname{var}_\delta(\hat\tau) &= \frac{1}{n_{rk}}\left(\frac{S_{tt}}{p_t^2} + \frac{S_{cc}}{p_c^2} + 2\frac{S_{tc}}{p_t p_c}\right) \\
&= \frac{p_r(1-p_r)}{n_{rk}^2}\left(\sum_{i\in\mathcal{R}_k} \frac{(Y_{it}-\mu_t)^2}{p_r^2} + \frac{(Y_{ic}-\mu_c)^2}{(1-p_r)^2} + 2\frac{(Y_{it}-\mu_t)(Y_{ic}-\mu_c)}{p_r(1-p_r)}\right) \\
&= \frac{p_r(1-p_r)}{n_{rk}^2}\sum_{i\in\mathcal{R}_k}\left(\frac{(Y_{it}-\mu_t)(1-p_r) + (Y_{ic}-\mu_c)p_r}{p_r(1-p_r)}\right)^2 \\
&= \frac{\bar\sigma_{rk}^2}{p_r(1-p_r)n_{rk}}
\end{aligned}$$

where $\bar\sigma_{rk}^2 = (1/n_{rk})\sum_{i\in\mathcal{R}_k}[(Y_{it}-\mu_t)(1-p_r) + (Y_{ic}-\mu_c)p_r]^2$. Under Assumption 4, $Y_{it} - \mu_t = Y_{ic} - \mu_c$ and $\bar\sigma_{rk}^2$ simplifies as given. Similarly, substituting $p_r = 1/2$ yields the other given simplification.