eAppendix

eAppendix A: Algorithms

In this section, we present two approaches for profile matching: one as a multidimensional knapsack problem and as a modified version of cardinality matching.[1]

A1: Multidimensional knapsack formulation

Profile matching can be expressed as a multidimensional knapsack problem[2] that directly finds the largest self-weighted balanced sample for each treatment group. Formally, let t index the units in treatment group $\mathcal{T}_T$. Define $\boldsymbol{X}_t$ as the vector of observed covariates of unit t. Denote $B_k(\boldsymbol{X}_t)$ as the k-th transformation of the covariates with $k = 1, ..., K$. Write $\boldsymbol{x}^*$ for the profile of vector covariate values of an individual or population. The multidimensional knapsack problem formulation is:

$$\begin{aligned} &\text{maximize}_{\boldsymbol{m}_t} && \sum_{t \in \mathcal{T}_T} m_t \\ &\text{subject to} && \left| \sum_{t \in \mathcal{T}_T} m_t B_k(\boldsymbol{X}_t) - m_t \boldsymbol{x}^* \right| \leq \sum_{t \in \mathcal{T}_T} m_t \delta_k, k = 1, ..., K \\ & && m_t \in \{0, 1\}, \forall t \in \mathcal{T}_T. \end{aligned} \tag{1}$$

Where m_t are matching or unit selector indicators, and δ_k is an imbalance tolerance defined by the investigator. The constraint imposes balance relative to the covariate profile defined by $B_k(\boldsymbol{X}^*)$. Therefore, the objective function maximizes the size of the sample that is balanced relative to the profile. By discarding the copies and keeping the original matched units Equation 1 obtains the largest self-weighted sample that satisfies the covariate balance requirements without a fixed matching ratio.

A2: Cardinality matching formulation

Profile matching can also be reexpressed in such a way as to make clear its connection with existing matching methods—namely, cardinality matching.[1] This makes it readily implementable with existing software in the `designmatch` package for R. Under this reexpression, for each treatment group $\mathcal{T}_T$, create a copy $\tilde{\mathcal{T}}_T$ of identical units and solve the optimization

problem:

$$
\begin{aligned}
\text{maximize}_{\boldsymbol{m}_t, \tilde{\boldsymbol{m}}_{\tilde{t}}} \quad & \sum_{t \in \mathcal{T}_T} m_t \\
\text{subject to} \quad & \sum_{t \in \mathcal{T}_T} m_t = \sum_{\tilde{t} \in \tilde{\mathcal{T}}_T} \tilde{m}_{\tilde{t}} \\
& \left| \sum_{t \in \mathcal{T}_T} m_t B_k(\boldsymbol{X}_t) - m_t \boldsymbol{x}^* \right| \leq \sum_{t \in \mathcal{T}_T} m_t \delta_k, k = 1, ..., K \qquad (2) \\
& \left| \sum_{\tilde{t} \in \tilde{\mathcal{T}}_T} m_{\tilde{t}} B_k(\boldsymbol{X}_{\tilde{t}}) - m_{\tilde{t}} \boldsymbol{x}^* \right| \leq \sum_{\tilde{t} \in \tilde{\mathcal{T}}_T} \tilde{m}_{\tilde{t}} \delta_k, k = 1, ..., K \\
& m_t \in \{0, 1\}, \forall t \in \mathcal{T}_T \\
& \tilde{m}_{\tilde{t}} \in \{0, 1\}, \forall \tilde{t} \in \tilde{\mathcal{T}}_T
\end{aligned}
$$

In the above, the new first constraint requires the same number of treated units to be selected from the original treatment group $\mathcal{T}_T$ as from its copy $\tilde{\mathcal{T}}_T$.

eAppendix B: R Code

In this section, we provide instructions on how to implement profile matching using existing cardinality matching software with the `designmatch` package for `R`. First, however, we recommend installing `gurobi`, an optimizer which increases the performance of `designmatch`. Instructions on installation can be found on `https://www.gurobi.com`.

Next, install the `designmatch` package in `R` via the code `install.packages("designmatch")`. Now, in this section, we provide example code of how to use `profmatch` to balance two treatment groups relative to a covariate profile calculated from the data. Note that the user can also supply their own covariate profile (e.g., from an external data source).

First, we load the `designmatch` package and read in the data.

```
> library(designmatch)
data("lalonde", package = "designmatch")
```

Next, we specify the covariates to be balanced.

```
> covs = c("age", "education", "black", "hispanic", "married", "nodegree", "re74",
"re75")
```

For each covariate, we specify the value to balance toward (which together form the profile).

In this case, the profile comprises the overall sample means of the covariates.

```
> mom_targets = colMeans(lalonde[, covs])
```

Additionally, we specify the imbalance tolerances. That is, `profmatch` will select units from each treatment group so that each matched group differs at most by `mom_tols` from the respective moments in `mom_targets`.

```
> cov_sds = apply(lalonde[, covs], 2, sd)
> mom_tols = 0.05 * cov_sds
```

Next, we specify some arguments required by the profmatch function.

```
> ## Vector of treatment group indicators
> t_ind = lalonde$treatment
>
> ## Covariate matrix
> mom_covs = as.matrix(lalonde[, covs])
>
> ## Putting it all together
> mom = list(covs = mom_covs, tols = mom_tols, targets = mom_targets)
```

Now we specify the solver options. Here, we use Gurobi, which is recommended

```
> t_max = 60*30
> solver = "gurobi"
> approximate = 0
> solver = list(name = solver, t_max = t_max, approximate = approximate, round_cplex
= 0, trace = 0)
```

Now, we are ready to profile match.

```
> ## Performing profile matching
> pmatch_out = profmatch(t_ind, mom, solver)
>
> ## Selecting the units that are matched
> lalonde_matched = lalonde[pmatch_out$id,]
```

Appendix C: Additional data details for the NSDUH case study

Data for this paper come from four consecutive years (2015–2018) of the National Survey on Drug Use and Health (NSDUH), resulting in a total sample size of 171,766 individuals. The NSDUH is administered annually to provide nationally representative information on tobacco, alcohol, and drug use, as well as on mental health and other health issues. We use this cross-sectional data set primarily for illustrative purposes—that is, to demonstrate the matching methods, rather than for the purposes of devising a precise portrait of effect estimates (see Samples et al.[3] for an extensive analysis using the first three years of this data set). For different targets, corresponding to a given subgroup, a particular population, and a specific individual, we are interested in understanding the relationship between opioid use and psychological distress and suicidal thoughts or behaviors, after adjusting for differences in subjects' observed characteristics. As long as the covariates, the exposure, and the outcomes are measured orderly in time and are subject to the usual assumptions for identification and generalization or transportability in observational studies (see, e.g., Dahabreh et al.)),[4] our estimates can be granted a causal interpretation. In what follows, we describe these three sets of variables.

The exposure—opioid use in the past 12 months—takes on one of three values for each individual: 0 = no opioid use in the past 12 months, 1 = opioid use but no misuse in the past 12 months, and 2 = opioid misuse in the past 12 months. We examine two outcomes related to this exposure: one continuous and one binary. The continuous outcome ranges from 0 to 24 and indicates the respondent's level of psychological distress over the past 30 days (a higher score indicates more distress). This variable is based on data collected from a series of six questions asking about symptoms of psychological distress over the past 30 days, such as feeling hopeless, nervous, restless or fidgety, and sad or depressed. The binary outcome (1 = yes, 0 = no) indicates whether the respondent has either seriously thought

about suicide, made plans to kill themselves, or attempted to kill themselves in the past 12 months.

Covariates used for matching include those associated with opioid use or misuse and suicidal behaviors in prior studies and sociodemographic measures.[3] Past-year health and behavioral measures include type of healthcare coverage (Medicare, Medicaid/SCHIP, military health insurance, private health insurance, group health insurance, or none); use of specific drugs (heroin, cocaine, methamphetamines, stimulants, and tranquilizers or sedatives); nicotine dependence (based on the Nicotine Dependence Syndrome Scale and the Fagerström Dependence Syndrome Scale); depression, alcohol, or marijuana use disorder (each based on criteria from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition); and self-rated health. Lifetime health and behavioral measures include number of chronic health conditions (diabetes, chronic bronchitis or COPD, cirrhosis, Hepatitis B or C, kidney disease, asthma, HIV/AIDS, cancer, and high blood pressure). Sociodemographic measures include age, sex, race/ethnicity, education, current employment status, income, marital status, number of children, and urban residence status.

For the generalization application, we define a subgroup of respondents who indicated they are a sexual minority. We select this subgroup in response to increased calls for more investigation into the relationship between opioid use and suicide for this group.[5] Survey respondents were asked whether they identified as one of the following: (1) heterosexual, (2) lesbian or gay, or (3) bisexual. We classify those who identified as lesbian or gay or bisexual as a sexual minority for the sake of our analysis. Some individuals (3,206, or 1.8 percent of respondents) responded to this question with a response of "don't know" or "refuse" or by leaving it blank. To avoid losing these sample members in the analysis, we reference another question from the survey. This question asked respondents what sexes they are attracted to. Respondents who indicated any attraction to the same sex are also classified as sexual minorities, if their response to the earlier question was "don't know" or "refuse,"

or if they left it blank. This results in an additional 1,037 individuals (0.6 percent) able to be classified.

For the transportability application, we construct matched samples whose covariate means resemble those of the Appalachian United States, a region that has suffered particularly severely from the recent opioid epidemic.[6] We use external data reports to define the covariate profile toward which we seek to balance the exposure groups.[7,8] We match on covariates from the previous list that are also included in these reports. To supplement cases where the desired covariate is not available across data sources, we use related and available covariates. Specifically, the following covariates are unavailable from the data sources on the Appalachian population: type of healthcare coverage (other than no healthcare coverage); use of specific drugs; nicotine dependence; depression, alcohol, or marijuana use disorder; self-rated health; number of chronic health conditions; current employment status; income; marital status, and number of children. Additionally, we modify the age and education categories to align across data sources. We supplement the excluded economic covariates with a measure of poverty and the excluded substance and alcohol use covariates with a measure of excessive alcohol use (measured according to criterion specified by the Centers for Disease Control and Prevension), which are available in our data and in the external reports. We also include veteran status as a covariate, as this is available across data sources.

For the personalization application, we construct matched samples whose covariate means resemble the characteristics of a hypothetical vulnerable patient. We define our vulnerable patient as a rural White male between 26 and 34 years old with a high school education whose past year psychological distress score indicates severe impairment. This last covariate was measured in the same way as the continuous outcome, albeit for the most difficult month in the prior 12 months, providing a score where "severe" was defined as 13 or above. Our patient has a score of 13. This covariate was measured for some NSDUH respondents and not others; respondents whose most difficult month in the past year was the current month

(i.e., whose psychological distress covariate was the same as the outcome) are missing on this measure, and so we exclude them from the matching procedure.

eAppendix D: Alternative implementation of profile matching

D1: Covariate Balance

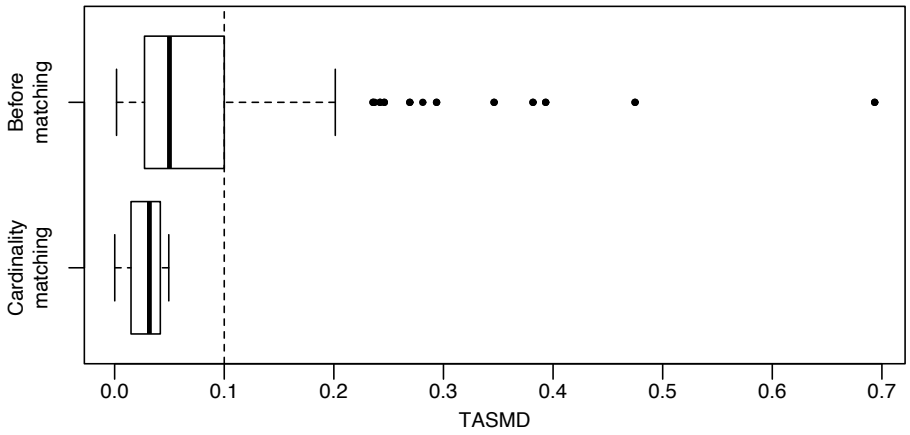
As an alternative method of matching—and one that preserves the 1:1 matching between units—we describe a method of pairwise matching of treatment groups, where the groups are simultaneously balanced toward the profile and matched. This falls under the method of cardinality matching. Thus, we have a total of six matched sets, representing all possible pairwise matchings across the three exposure groups (no opioid use, opioid use without misuse, and opioid misuse).

For the sexual minority target population, we construct the matched sets (a no opioid use-opioid misuse matched set and an opioid use without misuse-opioid misuse matched set) by balancing the exposure groups toward the distribution of covariates in the overall sexual minority sample (i.e., not within a particular exposure group) to estimate an average association, rather than an association within one exposure group. The no opioid use-opioid misuse cardinality matched set has 812 individuals ($812/2 = 406$ for each exposure group) and the opioid use without misuse-opioid misuse cardinality matched set has 832 individuals ($832/2 = 416$ for each exposure group). Of note, for this matching problem, the maximal achievable sample size of any one matched set (i.e., if we performed 1:1 matching without replacement and did not discard individuals, thereby ignoring the balance restriction) is 2,514. Because a greater number of sexual minorities in our data have not used opioids or used opioids without misusing them (3,753 and 6,614, respectively) as opposed to having misused them (1,257), the maximal achievable sample size of a matched set (which includes two treatment groups) is equal to twice the number of sexual minorities who have misused

opioids (2 x 1,257 = 2,514), reflecting our use of 1:1 matching without replacement.

We use the target absolute standardized mean difference (TASMD)[9] to summarize the performance of this method in terms of its ability to achieve covariate balance relative to the target population. The TASMD measures the standardized difference between the mean in one population and the mean in the target population. In Figure 1, we see that profile matching achieves good balance relative to the target population.

Figure 1: Covariate Balance Relative to the Sexual Minorities Target Population Before and After Cardinality Matching, National Survey on Drug Use and Health, 2015-2018

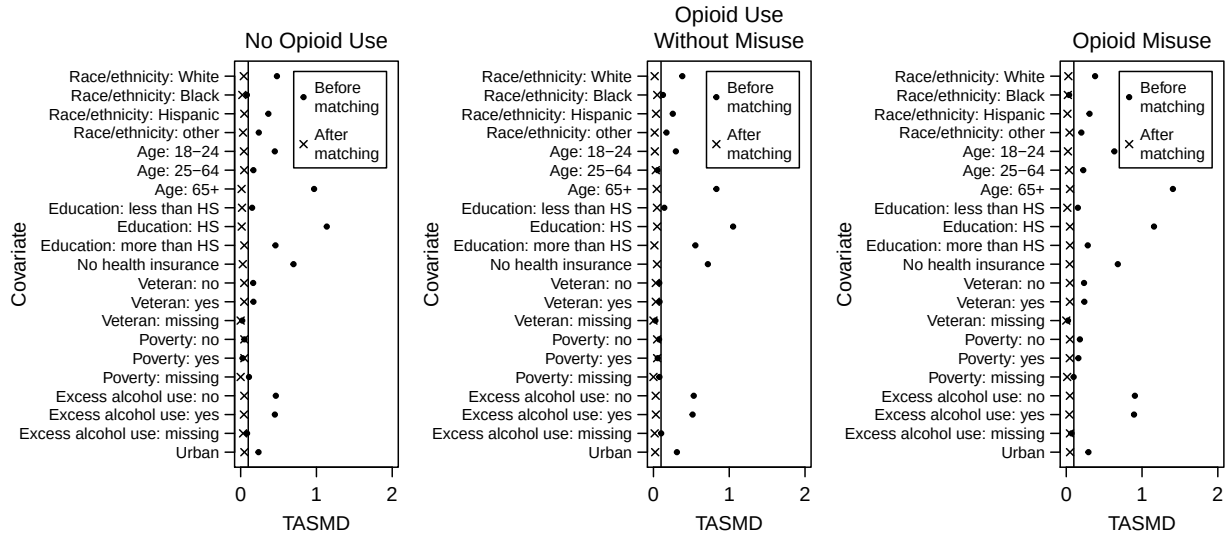


The target average standardized mean differences (TASMDs) are plotted before matching, after propensity score matching, and after cardinality matching for balance toward the sexual minorities subgroup. The TASMD measures, for each covariate, its difference—appropriately standardized by the pooled standard deviation across the three treatment groups (no opioid use, opioid use without misuse, and opioid misuse in the past year) in the study sample—from the value in the target population: the subgroup of sexual minorities. The samples are from the National Survey on Drug Use and Health 2015-2018.

For the Appalachian target population, the no opioid use-opioid misuse cardinality matched set has 2,224 individuals, with 1,112 from each exposure group. The opioid use without misuse-opioid misuse cardinality matched set has 2,208 individuals, with 1,104 from each exposure group. Because this population is defined in terms of a distribution of covariates in

an out-of-sample population, there is no notion of "maximal achievable sample size" as with the sexual minority cardinality matched sets. We summarize the performance of cardinality matching in constructing matched samples whose covariate distribution resembles that of a target population in Figure 2, which plots the TASMDs for each covariate before and after cardinality matching, for each exposure group. Note that, as described above, a cardinality matched data set for the exposure group defined by opioid misuse is actually constructed twice—one for matching with the no opioid use exposure group and one for matching with the opioid use without misuse exposure group. For simplicity, we display the results for the smaller group (i.e., the opioid misuse group constructed for the opioid use without misuse-opioid misuse matched set), as the results are fairly similar for the two opioid misuse cardinality matched sets. Overall, Figure 2 shows that groups that were relatively imbalanced relative to the target population could be well-balanced using cardinality matching.

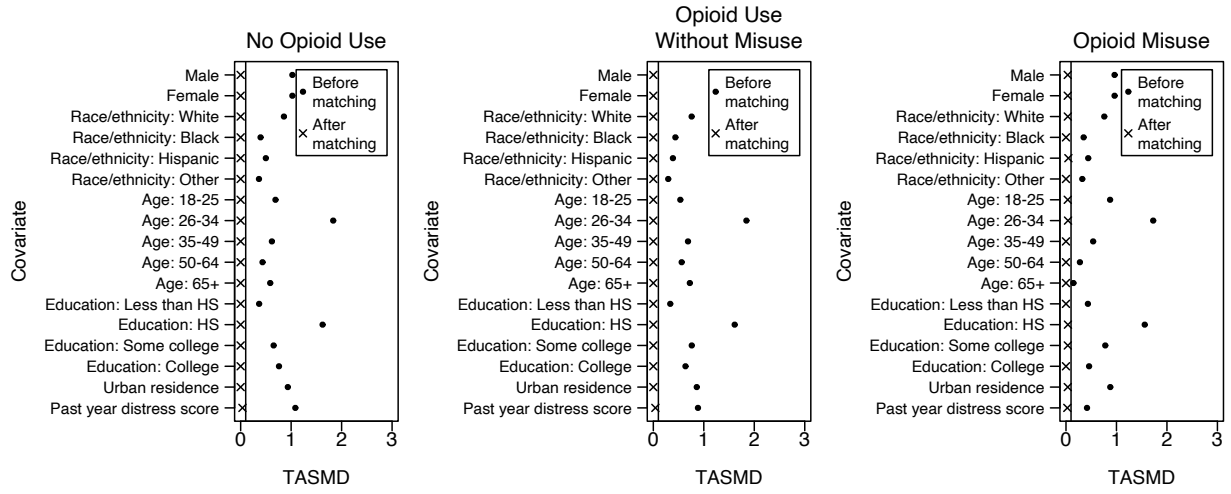
Figure 2: Covariate Balance Relative to the Appalachian Target Population Before and After Cardinality Matching, National Survey on Drug Use and Health, 2015-2018



The target average standardized mean differences (TASMDs) are plotted before and after cardinality matching for balance toward the Appalachian target population. The TASMD measures, for each covariate, its difference—appropriately standardized by the pooled standard deviation across the three treatment groups in the study sample—from the value in the Appalachian target population. The matched samples are from the National Survey on Drug Use and Health, 2015-2018, and the covariate information for the Appalachian target population comes from the 2012-2016 American Community Survey. HS = high school.

For the vulnerable patient profile, the no opioid use-opioid misuse cardinality matched set has 114 individuals, with 57 from each exposure group. The opioid use without misuse-opioid misuse cardinality matched set has 112 individuals, with 56 from each exposure group. Cardinality matching achieved good balance relative to the profile of the vulnerable patient for each exposure group, with TASMDs below 0.1 for all covariates included in the matching procedure (see Figure 3).

Figure 3: Covariate Balance Relative to the Vulnerable Patient Before and After Cardinality Matching, National Survey on Drug Use and Health, 2015-2018



The target average standardized mean differences (TASMDs) are plotted before and after cardinality matching for balance toward the characteristics of a vulnerable patient defined as as a rural White male between 26 and 34 years old with a high school education whose past year psychological distress score indicates severe impairment. The TASMD measures, for each covariate, its difference—appropriately standardized by the pooled standard deviation across the three treatment groups (no opioid use, opioid use without misuse, and opioid misuse in the past year) in the study sample—from the vulnerable patient's value. The matched samples and the profile for the vulnerable patient are from the National Survey on Drug Use and Health 2015-2018. HS = high school.

After using cardinality matching to find the largest pairwise matched samples that are balanced, we re-match the samples to minimize the covariate distances between the matched units. That is—cardinality matching produces equally-sized subsamples from two exposure groups, and re-matching allows us to find pairs of individuals (i.e., one from each exposure group) such that the overall sum of covariate distances across the matched pairs is minimized.[10] Thus, the final re-matched sets consist of the same individuals—they are just paired, perhaps, differently. This re-matching reduces heterogeneity in the matched pairs and translates into less sensitivity to hidden bias if the covariates used to compute the

distances are strong predictors of the outcome.[1,10] To achieve this property, we construct matched pairs based on the binary covariate indicating the occurrence of a past year major depressive episode, as this covariate is predictive of both outcomes.

D2: Outcome Analysis

To measure the relationship between the opioid misuse exposure (relative to either no opioid use or to opioid use without misuse) and the binary outcome of past year suicidal thoughts, plans, or attempts; we compute the average difference in the binary outcome for the matched pairs. McNemar's test for paired binary data is used to compute the statistical significance of each association.[11] Table 1 presents these results. Generally, past year opioid misuse is associated with an increased probability of past year suicidal thoughts, plans, or attempts—regardless of whether the increased probability is relative to no opioid use or relative to opioid use without misuse. In half of the comparisons, the association does not reach statistical significance. For the vulnerable patient associations, it is relevant to note that the sample sizes are relatively small, with 56 or 57 individuals in each exposure group.

To measure the relationship between the opioid misuse exposure (relative to either no opioid use or to opioid use without misuse) and the continuous outcome of past month psychological distress, we compute the average difference in the outcome for the matched pairs. Wilcoxon's signed rank test for paired continuous outcomes is used to test for statistical significance.[11] These results are presented in Table 2. Generally, past year opioid misuse is associated with higher levels of psychological distress—regardless of whether the increase in distress is relative to no opioid use or relative to opioid use without misuse. In the comparisons for the vulnerable patient, the association does not reach statistical significance, and again, it is relevant to note the small sample sizes for these matched sets.

To aid interpretation of results, Figure 4 presents the distributions of pairwise differences in the continuous outcome for each estimand. If most of the boxplot and points for a given association appear above the horizontal dashed line at 0, there is evidence of a positive

Table 1: Associations Between Opioid Use Types and Past Year Suicidal Thoughts, Plans, or Attempts, National Survey on Drug Use and Health, 2015-2018

	No opioid use-opioid misuse matched set				Opioid use without misuse-opioid misuse matched set			
	Matched sample mean				Matched sample mean			
Cardinality matched sample	Opioid misuse	No opioid use	Difference	P value	Opioid misuse	Opioid use without misuse	Difference	P value
Sexual minorities	0.26	0.21	0.05	0.11	0.25	0.18	0.07	0.03
Appalachian target population	0.14	0.04	0.09	<0.01	0.15	0.07	0.08	<0.01
Vulnerable patient profile	0.23	0.12	0.11	0.11	0.21	0.18	0.04	0.79

The probabilities of past year suicidal thoughts, behaviors or attempts are displayed by treatment group (no opioid use, opioid use without misuse, and opioid misuse in the past year) for the three targets: the sexual minorities subgroup, the Appalachian target population, and the vulnerable patient. Results reflect cardinality matching each pair of treatment groups for balance toward each other and toward each target. Two-sided P values are based on McNemar's test for paired binary data. Data come from the National Survey on Drug Use and Health, 2015-2018.

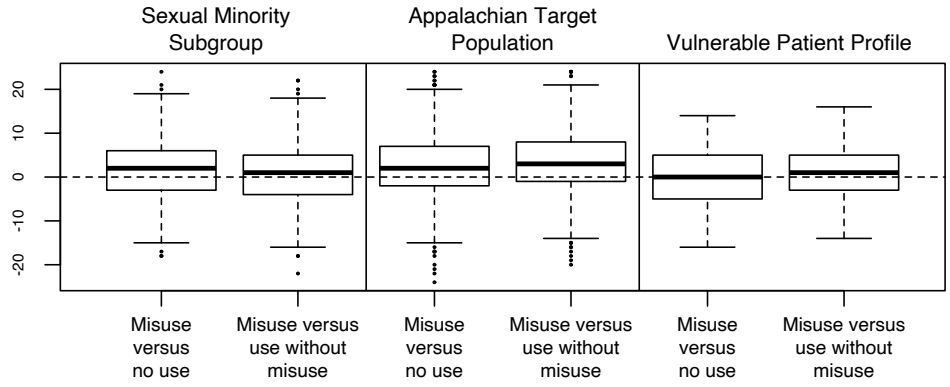
Table 2: Associations Between Opioid Use Types and Past Month Psychological Distress, National Survey on Drug Use and Health 2015-2018

	No opioid use-opioid misuse matched set				Opioid use without misuse-opioid misuse matched set			
	Matched sample mean				Matched sample mean			
Cardinality matched sample	Opioid misuse	No opioid use	Difference	P value	Opioid misuse	Opioid use without misuse	Difference	P value
Sexual minorities	8.9	7.3	1.6	<0.01	8.8	8.0	0.7	0.03
Appalachian target population	7.0	3.5	3.5	<0.01	7.2	4.8	2.4	<0.01
Vulnerable patient profile	9.2	8.1	1.1	0.10	9.1	8.9	0.3	0.4

The means of past month psychological distress score for treatment group (no opioid use, opioid use without misuse, and opioid misuse in the past year) are compared pairwise for the three targets: the sexual minorities subgroup, the Appalachian target population, and the vulnerable patient. Results reflect cardinality matching each pair of treatment groups for balance toward each other and toward each target. Two-sided P values are based on Wilcoxon's signed rank test for paired continuous outcomes. Data come from the National Survey on Drug Use and Health, 2015-2018.

association between opioid misuse and past month psychological distress. Similarly, if most of the results appear below the horizontal dashed line, there is evidence of a negative association. Generally, the associations are modestly positive.

Figure 4: Distributions of Cardinality Matched Pair Differences in Past Month Psychological Distress Score for Various Targets, National Survey on Drug Use and Health 2015-2018



Cardinality matched pairwise differences in past month psychological distress score are plotted after cardinality matching for balance toward three targets: the sexual minority subgroup, the Appalachian target population, and the vulnerable patient. Results are shown as boxplots. Data are from the National Survey on Drug Use and Health 2015-2018.

D3: Sensitivity Analysis

We test the sensitivity of statistically significant associations by testing how much hidden bias would need to be present in order to meaningfully alter conclusions. We implement the sensitivity analysis of Rosenbaum,[11] which we briefly describe below.

Let π_{ij} denote the probability that unit j in matched pair i misuses opioids. If there is no hidden bias, then the probability of opioid misuse for two units in a matched pair (e.g., unit j and unit j' in matched pair i) is the same. If there is hidden bias, then their odds of opioid misuse may differ. Suppose that their odds of opioid misuse differ by at most a factor of Γ.

That is:

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}/(1-\pi_{ij})}{\pi_{ij'}/(1-\pi_{ij'})} \leq \Gamma$$

If $\Gamma = 1$, there is no hidden bias. Rosenbaum's sensitivity test finds the largest value of Γ where we still reject the null hypothesis of an association (i.e., $P < 0.05$). We sought such a Γ for each statistically significant result above. These results are presented in Table 3. Generally, the selected results for sexual minorities are more sensitive to hidden bias than are those for the Appalachian target population, and the results are sensitive for a range of gammas—some very low to some relatively high—from 1.02 to 2.80. To guide interpretation of the results, consider the second to last row of Table 3. Here, $\Gamma = 2.80$, meaning that matched pairs of those not using opioids and those misusing opioids can differ in their odds of opioid misuse by 180 percent without altering the conclusion about the association between opioid misuse (relative to no opioid use) and past month psychological distress.

Table 3: Sensitivity of Significant Associations to Hidden Biases, National Survey on Drug Use and Health, 2015-2018

Association	Γ
Suicidal thoughts, plans, or attempts and opioid misuse relative to...	
Opioid use without misuse, for sexual minorities	1.30
No opioid use, for the Appalachian target population	2.76
Opioid use without misuse, for the Appalachian target population	1.61
Psychological distress and opioid misuse relative to...	
No opioid use, for sexual minorities	1.33
Opioid use without misuse, for sexual minorities	1.02
No opioid use, for the Appalachian target population	2.80
Opioid use without misuse, for the Appalachian target population	1.92

Results for Rosenbaum's sensitivity analysis are shown for the statistically significant associations from Table 1 and Table 2. Rosenbaum's bounds finds the largest value of Γ where we still reject the null hypothesis of an association, where $\Gamma = 1$ reflects no hidden bias. Data come from the National Survey on Drug Use and Health, 2015-2018.

eAppendix E: Additional simulation study results

In this section, we present additional results for the simulation study in Section 5. Table 6 presents the mean TASMDs across methods under no effect heterogeneity, some of which appear in the main text. Table 5 presents the mean absolute bias (MAB) and root mean square error (RMSE) results under no effect heterogeneity and high overlap. Table **??** presents the empirical coverage probability and average length of bootstrapped confidence intervals. under no effect heterogeneity and high overlap. Additional tables have results from the simulation study under the effect heterogeneity setting. Table 7 presents the mean TASMDs across methods under effect heterogeneity, and Table 8 presents the effective sample sizes. Table 9 presents mean absolute bias (MAB) and root mean square error (RMSE) results, and Table 10 presents results on the the empirical coverage probability and average length of bootstrapped confidence intervals. Additionally, we present tables showing the mean bias and variance of the estimators for the simulation study in Section 5. Table 11 presents results for the no effect heterogeneity setting (A), and Table 12 presents results for the effect heterogeneity setting (B).

Table 4: Mean TASMDs across methods under no effect heterogeneity (A)

		High Overlap							Low Overlap						
Covariate	Group	Before	PM1	PM2	PM3	IOW1	IOW2	IOW3	Before	PM1	PM2	PM3	IOW1	IOW2	IOW3
X_1	Treated	0.635	0.047	0.352	0.643	0.056	0.234	0.648	1.013	0.046	0.593	1.049	0.129	0.435	1.065
	Control	0.634	0.047	0.352	0.641	0.056	0.236	0.646	1.011	0.047	0.593	1.049	0.125	0.438	1.065
X_2	Treated	0.594	0.048	0.416	0.602	0.055	0.327	0.606	0.941	0.048	0.716	0.974	0.125	0.602	0.989
	Control	0.599	0.048	0.414	0.601	0.055	0.327	0.605	0.938	0.048	0.718	0.974	0.122	0.603	0.989
X_3	Treated	0.597	0.048	0.049	0.603	0.056	0.048	0.607	0.940	0.048	0.049	0.975	0.123	0.084	0.990
	Control	0.592	0.048	0.049	0.602	0.056	0.046	0.606	0.943	0.048	0.049	0.976	0.121	0.087	0.991
X_4	Treated	0.252	0.048	0.266	0.048	0.047	0.272	0.031	0.375	0.047	0.459	0.049	0.088	0.477	0.033
	Control	0.248	0.048	0.268	0.048	0.045	0.273	0.031	0.381	0.047	0.456	0.049	0.088	0.475	0.033
X_5	Treated	0.096	0.043	0.048	0.044	0.054	0.058	0.039	0.135	0.043	0.060	0.047	0.099	0.096	0.045
	Control	0.096	0.043	0.048	0.044	0.053	0.058	0.039	0.136	0.043	0.059	0.047	0.098	0.097	0.045
X_6	Treated	0.084	0.039	0.039	0.039	0.045	0.039	0.031	0.112	0.037	0.039	0.043	0.090	0.058	0.034
	Control	0.072	0.038	0.039	0.039	0.046	0.039	0.031	0.111	0.037	0.039	0.043	0.089	0.058	0.033

IOW = inverse odds weighting; PM = profile matching. Accompanying numbers indicate the specification of each method.

Table 5: Mean Absolute Bias (MAB) and Root Mean Square Error (RMSE) across methods and settings under no effect heterogeneity (A) and high overlap

Outcome	Model	PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	MAB	0.09	0.21	0.12	0.27	0.24	0.19	0.08	0.21	0.11	0.08	0.22	0.11
	RMSE	0.11	0.26	0.15	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
OM 2	MAB	0.09	0.17	0.14	0.22	0.19	0.15	0.09	0.17	0.14	0.10	0.18	0.13
	RMSE	0.11	0.21	0.18	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
OM 3	MAB	0.74	0.60	0.71	1.10	1.46	1.36	0.75	0.58	0.60	1.53	0.81	0.74
	RMSE	0.93	0.76	0.90	0.01	0.04	<0.01	<0.01	<0.01	0.01	0.02	0.03	<0.01

PM = profile matching; aPM = profile matching augmented with an outcome model; IOW = inverse odds weighting; aIOW = inverse odds weighting augmented with an outcome model. Accompanying numbers indicate the specification of each method.

Table 6: Coverage probability (cov.) and length (len.) of bootstrapped confidence intervals across methods and settings under no effect heterogeneity (A) and high overlap

Outcome Model		PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	Cov.	0.97	0.96	0.96	0.95	0.95	0.95	0.96	0.96	0.96	0.95	0.95	0.96
	Len.	0.46	1.06	0.64	1.31	1.24	0.96	0.40	1.05	0.60	0.40	1.14	0.56
OM 2	Cov.	0.96	0.96	0.96	0.95	0.95	0.95	0.96	0.96	0.96	0.95	0.95	0.95
	Len.	0.45	0.88	0.73	1.08	0.95	0.73	0.43	0.88	0.72	0.46	0.89	0.65
OM 3	Cov.	0.96	0.95	0.95	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95
	Len.	3.74	3.09	3.49	5.30	7.77	6.63	3.76	2.84	2.96	7.10	4.00	3.51

PM = profile matching; aPM = profile matching augmented with an outcome model; IOW = inverse odds weighting; aIOW = inverse odds weighting augmented with an outcome model Accompanying numbers indicate the specification of each method.

Table 7: Mean TASMDs across methods under effect heterogeneity (B)

		High Overlap							Low Overlap						
Covariate	Group	Before	PM1	PM2	PM3	IOW1	IOW2	IOW3	Before	PM1	PM2	PM3	IOW1	IOW2	IOW3
X_1	Treated	0.635	0.047	0.352	0.643	0.058	0.234	0.648	1.013	0.047	0.590	1.048	0.125	0.436	1.064
	Control	0.634	0.047	0.352	0.642	0.057	0.234	0.647	1.011	0.046	0.591	1.047	0.126	0.436	1.064
X_2	Treated	0.594	0.048	0.416	0.603	0.057	0.328	0.608	0.941	0.048	0.715	0.976	0.120	0.601	0.991
	Control	0.599	0.048	0.415	0.602	0.056	0.327	0.607	0.938	0.048	0.715	0.972	0.120	0.602	0.988
X_3	Treated	0.597	0.048	0.049	0.604	0.057	0.047	0.608	0.940	0.048	0.049	0.975	0.121	0.085	0.990
	Control	0.592	0.048	0.049	0.603	0.056	0.047	0.607	0.943	0.048	0.049	0.974	0.120	0.085	0.990
X_4	Treated	0.252	0.048	0.268	0.048	0.045	0.274	0.031	0.375	0.047	0.464	0.049	0.088	0.481	0.034
	Control	0.248	0.048	0.268	0.048	0.045	0.274	0.031	0.381	0.047	0.462	0.049	0.087	0.481	0.034
X_5	Treated	0.096	0.043	0.048	0.044	0.052	0.059	0.038	0.135	0.043	0.059	0.047	0.097	0.095	0.046
	Control	0.096	0.043	0.048	0.044	0.052	0.058	0.038	0.136	0.043	0.058	0.047	0.098	0.096	0.046
X_6	Treated	0.084	0.039	0.039	0.040	0.045	0.039	0.031	0.112	0.037	0.039	0.043	0.089	0.059	0.034
	Control	0.072	0.038	0.038	0.039	0.046	0.038	0.031	0.111	0.037	0.039	0.043	0.089	0.059	0.035

IOW = inverse odds weighting; PM = profile matching. Accompanying numbers indicate the specification of each method.

Table 8: Mean effective sample size across methods under effect heterogeneity (B)

	PM1	PM2	PM3	IOW1	IOW2	IOW3
High Overlap	414.0	468.3	637.2	401.9	504.1	682.5
Low Overlap	194.3	289.4	572.8	149.1	281.1	614.9

IOW = inverse odds weighting; PM = profile matching. Accompanying numbers indicate the specification of each method.

Table 9: Mean Absolute Bias (MAB) and Root Mean Square Error (RMSE) across methods and settings under effect heterogeneity (B)

		High Overlap											
Outcome Model		PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	MAB	0.18	0.26	0.19	0.43	0.39	0.34	0.15	0.24	0.17	0.22	0.30	0.24
	RMSE	0.23	0.32	0.24	0.18	0.19	0.18	0.02	0.03	0.02	0.18	0.18	0.18
OM 2	MAB	0.17	0.22	0.20	0.38	0.34	0.30	0.16	0.21	0.19	0.23	0.27	0.25
	RMSE	0.22	0.28	0.25	0.17	0.19	0.18	0.02	0.03	0.02	0.18	0.18	0.18
OM 3	MAB	0.76	0.61	0.74	1.16	1.51	1.43	0.77	0.59	0.62	1.54	0.83	0.79
	RMSE	0.95	0.77	0.93	0.22	0.19	0.20	0.07	0.03	0.02	0.24	0.17	0.18
		Low Overlap											
Outcome Model		PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	MAB	0.21	0.30	0.19	0.73	0.55	0.39	0.20	0.29	0.18	0.31	0.41	0.30
	RMSE	0.27	0.37	0.24	0.26	0.25	0.28	0.10	0.09	0.07	0.26	0.26	0.27
OM 2	MAB	0.21	0.26	0.20	0.64	0.46	0.35	0.21	0.25	0.19	0.33	0.37	0.30
	RMSE	0.27	0.32	0.25	0.26	0.26	0.27	0.10	0.09	0.07	0.27	0.26	0.27
OM 3	MAB	0.98	0.72	0.72	1.85	2.54	1.80	0.98	0.70	0.62	3.59	1.26	0.88
	RMSE	1.26	0.91	0.91	0.28	0.25	0.22	0.10	0.09	0.07	0.29	0.30	0.24

PM = profile matching; aPM = profile matching augmented with an outcome model; IOW = inverse odds weighting; aIOW = inverse odds weighting augmented with an outcome model. Accompanying numbers indicate the specification of each method.

Table 10: Coverage probability (cov.) and length (len.) of bootstrapped confidence intervals across methods and settings under effect heterogeneity (B)

		High Overlap											
Outcome Model		PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	Cov.	0.96	0.96	0.96	0.94	0.94	9.92	0.96	0.96	0.95	0.86	0.92	0.89
	Length	0.90	1.31	0.97	1.97	1.82	1.51	0.77	1.23	0.86	0.82	1.36	0.91
OM 2	Cov.	0.95	0.96	0.96	0.94	0.93	0.92	0.95	0.96	0.95	0.87	0.91	0.89
	Length	0.86	1.14	1.01	1.72	1.54	1.27	0.78	1.09	0.95	0.85	1.14	0.97
OM 3	Cov.	0.96	0.96	0.95	0.95	0.96	0.95	0.96	0.95	0.95	0.95	0.94	0.94
	Length	3.81	3.18	3.58	5.48	8.05	6.85	3.81	2.92	3.04	7.14	4.07	3.65
		Low Overlap											
Outcome Model		PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	Cov.	0.94	0.95	0.95	0.93	0.93	0.89	0.93	0.95	0.95	0.80	0.91	0.78
	Length	1.01	1.49	0.95	3.13	2.46	1.58	0.93	1.44	0.87	1.01	2.41	0.89
OM 2	Cov.	0.94	0.94	0.95	0.92	0.92	0.86	0.94	0.95	0.95	0.83	0.88	0.79
	Length	1.00	1.27	0.97	2.72	2.00	1.29	0.95	1.23	0.91	1.15	1.76	0.91
OM 3	Cov.	0.96	0.96	0.95	0.94	0.97	0.95	0.96	0.95	0.95	0.95	0.94	0.94
	Length	4.95	3.68	3.59	7.78	12.74	8.67	4.96	3.47	3.08	15.98	8.35	4.04

PM = profile matching; aPM = profile matching augmented with an outcome model; IOW = inverse odds weighting; aIOW = inverse odds weighting augmented with an outcome model. Accompanying numbers indicate the specification of each method.

Table 11: Mean Bias and Variance across methods and settings under no effect heterogeneity (A)

		High Overlap											
Outcome Model		PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	Mean Bias	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|
	Variance	0.01	0.07	0.02	0.12	0.10	0.06	<0.01	0.07	0.02	0.01	0.08	0.02
OM 2	Mean Bias	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|
	Variance	0.01	0.05	0.03	0.08	0.06	0.03	0.01	0.05	0.03	0.01	0.05	0.03
OM 3	Mean Bias	|<0.01|	|<0.01|	|<0.01|	-0.01	-0.04	|<0.01|	|<0.01|	|<0.01|	-0.01	-0.02	0.03	|<0.01|
	Variance	0.86	0.58	0.80	2.16	4.97	3.17	0.90	0.53	0.57	3.90	1.19	0.92
		Low Overlap											
Outcome Model		PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	Mean Bias	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|
	Variance	0.02	0.09	0.02	0.44	0.25	0.07	0.02	0.09	0.02	0.05	0.21	0.02
OM 2	Mean Bias	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|	|<0.01|
	Variance	0.02	0.06	0.03	0.32	0.12	0.03	0.02	0.06	0.02	0.09	0.13	0.02
OM 3	Mean Bias	|<0.01|	|<0.01|	-0.01	-0.06	-0.05	-0.04	|<0.01|	|<0.01|	-0.01	-0.09	-0.01	-0.03
	Variance	1.55	0.77	0.80	8.31	21.55	6.11	1.57	0.72	0.60	51.63	4.76	1.41

PM = profile matching; aPM = profile matching augmented with an outcome model; IOW = inverse odds weighting; aIOW = inverse odds weighting augmented with an outcome model. Accompanying numbers indicate the specification of each method.

Table 12: Mean Bias and Variance across methods and settings under effect heterogeneity (B)

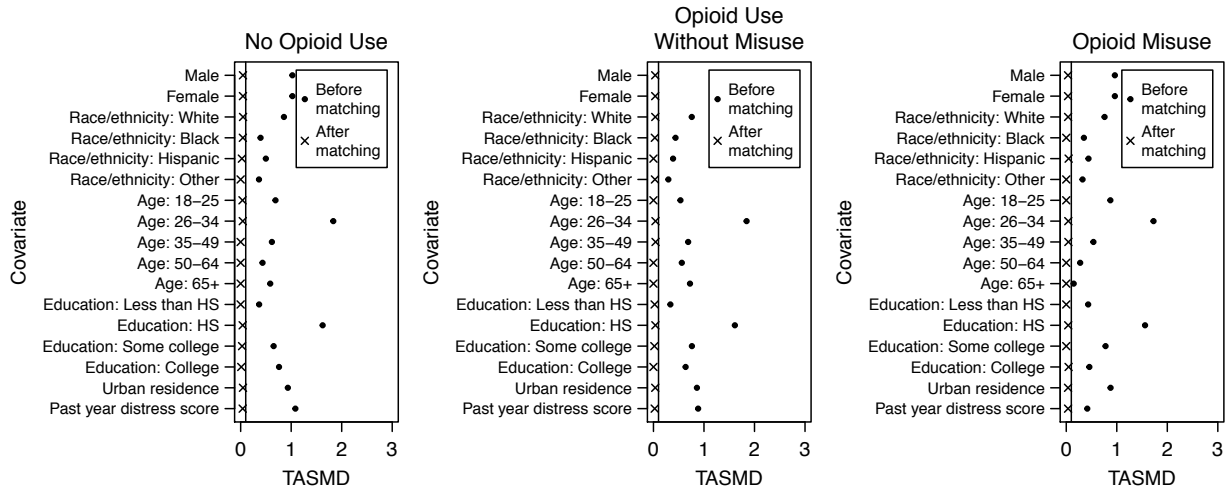
		High Overlap											
Outcome Model		PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	Mean Bias	-0.03	-0.03	-0.02	-0.18	-0.19	-0.18	-0.02	-0.03	-0.02	-0.18	-0.18	-0.18
	Variance	0.05	0.10	0.06	0.26	0.21	0.15	0.04	0.09	0.05	0.04	0.11	0.05
OM 2	Mean Bias	-0.03	-0.03	-0.02	-0.17	-0.19	-0.18	-0.02	-0.03	-0.02	-0.18	-0.18	-0.18
	Variance	0.05	0.08	0.06	0.20	0.15	0.10	0.04	0.07	0.05	0.05	0.08	0.06
OM 3	Mean Bias	-0.07	-0.03	-0.03	-0.22	-0.19	-0.20	-0.07	-0.03	-0.02	-0.24	-0.17	-0.18
	Variance	0.90	0.59	0.86	2.47	5.29	3.78	0.92	0.55	0.62	3.91	1.18	1.13
		Low Overlap											
Outcome Model		PM1	PM2	PM3	IOW1	IOW2	IOW3	aPM1	aPM2	aPM3	aIOW1	aIOW2	aIOW3
OM 1	Mean Bias	-0.11	-0.09	-0.07	-0.26	-0.25	-0.28	-0.10	-0.09	-0.07	-0.26	-0.26	-0.27
	Variance	0.06	0.13	0.05	0.88	0.47	0.16	0.05	0.12	0.05	0.08	0.21	0.05
OM 2	Mean Bias	-0.10	-0.09	-0.07	-0.26	-0.26	-0.27	-0.10	-0.09	-0.07	-0.27	-0.26	-0.27
	Variance	0.06	0.10	0.05	0.71	0.29	0.11	0.06	0.09	0.05	0.11	0.15	0.05
OM 3	Mean Bias	-0.10	-0.10	-0.07	-0.28	-0.25	-0.22	-0.10	-0.09	-0.07	-0.29	-0.30	-0.24
	Variance	1.59	0.82	0.82	8.13	22.68	5.95	1.60	0.76	0.61	45.60	3.53	1.35

PM = profile matching; aPM = profile matching augmented with an outcome model; IOW = inverse odds weighting; aIOW = inverse odds weighting augmented with an outcome model. Accompanying numbers indicate the specification of each method.

eAppendix F: Additional case study results

In this section, we summarize the performance of profile matching for constructing matched sets whose aggregate covariate distributions (in this case, means) resemble the profile of a hypothetical vulnerable patient.

Figure 5: Balance before and after profile matching toward the characteristics of a vulnerable patient



References

[1] José R. Zubizarreta, Ricardo D. Paredes, and Paul R. Rosenbaum. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics*, 8(1):204–231, 2014.

[2] Hans Kellerer, Ulrich Pferschy, and David Pisinger. *Knapsack Problems*. Springer-Verlag Berlin Heidelberg, 2004.

[3] Hillary Samples, Elizabeth A. Stuart, and Mark Olfson. Opioid use and misuse and suicidal behaviors in a nationally representative sample of us adults. *American Journal of Epidemiology*, 188(7):1245–1253, 2019.

[4] Issa J. Dahabreh, Sarah E. Robertson, Eric J. Tchetgen, Elizabeth A. Stuart, and Miguel A. Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

[5] Ehsan Moazen-Zadeh, Mohammad Karamouzian, Hannah Kia, Travis Salway, Olivier Ferlatte, and Rod Knight. A call for action on overdose among lgbtq people in north america. *The Lancet Psychiatry*, 6(9):725–726, 2019.

[6] Appalachian Regional Commission. Communicating about opioids in appalachia: challenges, opportunities, and best practices. Report, 2018a.

[7] Kelvin Pollard and Linda A. Jacobsen. The appalachian region: a data overview from the 2012-2016 american community survey. Report, Population Reference Bureau, 2018.

[8] Appalachian Regional Commission. Creating a culture of health in appalachia: disparities and bright spots. Report, 2018b.

[9] Ambarish Chattopadhyay, Christopher H Hase, and José R Zubizarreta. Balancing vs modeling approaches to weighting in practice. *Statistics in Medicine*, 39(24):3227–3254, 2020.

[10] Giancarlo Visconti and José R. Zubizarreta. Handling limited overlap in observational studies with cardinality matching. *Observational Studies*, 4:217–249, 2018.

[11] Paul R. Rosenbaum. *Observational Studies*. New York: Springer, 2002a.