Health Services and Outcomes Research Methodology
https://doi.org/10.1007/s10742-022-00280-0

A tutorial comparing different covariate balancing methods with an application evaluating the causal effects of substance use treatment programs for adolescents

Andreas Markoulidakis[1,2] · Khadijeh Taiyari[2] · Peter Holmans[3] · Philip Pallmann[2] · Monica Busse[2] · Mark D. Godley[4] · Beth Ann Griffin[5]

Received: 15 October 2021 / Revised: 12 April 2022 / Accepted: 14 May 2022

Abstract
Randomized controlled trials are the gold standard for measuring causal effects. However, they are often not always feasible, and causal treatment effects must be estimated from observational data. Observational studies do not allow robust conclusions about causal relationships unless statistical techniques account for the imbalance of pretreatment confounders across groups and key assumptions hold. Propensity score and balance weighting (PSBW) are useful techniques that aim to reduce the observed imbalances between treatment groups by weighting the groups to look alike on the observed confounders. Notably, there are many methods available to estimate PSBW. However, it is unclear a priori which will achieve the best trade-off between covariate balance and effective sample size for a given application. Moreover, it is critical to assess the validity of key assumptions required for robust estimation of the needed treatment effects, including the overlap and no unmeasured confounding assumptions. We present a step-by-step guide to the use of PSBW for estimation of causal treatment effects that includes steps on how to evaluate overlap before the analysis, obtain estimates of PSBW using multiple methods and select the optimal one, check for covariate balance on multiple metrics, and assess sensitivity of findings (both the estimated treatment effect and statistical significance) to unobserved confounding. We illustrate the key steps using a case study examining the relative effectiveness of substance use treatment programs and provide a user-friendly Shiny application that can implement the proposed steps for any application with binary treatments.

Keywords Propensity score · Balancing weights · Sensitivity analysis · Causal treatment effect · Unmeasured confounding

✉ Andreas Markoulidakis
MarkoulidakisA@cardiff.ac.uk

Extended author information available on the last page of the article

Published online: 27 May 2022

1 Introduction

In a randomized controlled trial (RCT), the assignment to treatment or control group is done using randomization to ensure that there is no systematic bias from observed and unobserved confounders when estimating the effect of the treatment. For sufficiently large sample sizes, the two groups will usually have similar baseline characteristics so that the groups are comparable to one another Altman and Bland (1999).

Unfortunately, such is not the case with observational studies, where randomization to treatment assignments is not possible. In observational studies, group assignment might be due to factors that the researcher cannot control, such as underlying conditions of the individual. These differences between the groups will introduce bias into the estimated effects of a treatment. For example, the choice to exercise among individuals with a disease might be influenced in part based on the severity of their disease: those with more severe disease might be less keen or able to exercise, and any subsequent comparison between those who did and did not exercise will be distorted, as it will compare groups with different pre-treatment levels of disease severity.

Relatedly, in many health services research applications, there may not be funding available to test the relative effectiveness of every possible evidence-based treatment or health service via RCTs. Instead, health services researchers are left with access to rich, secondary data sets (like electronic health records, insurance data sets, or cohort studies) that follow individuals over time who self-select into different types of treatments or health services. Naturally, the group of individuals who self-select into treatments or health services will be notably different from those that do not in terms of both sociodemographic characteristics and other health indicators. This type of confounding must be removed before estimating treatment effects. Our motivating example is based on such a rich data source, the Global Appraisal of Individual Needs (GAIN) data base which has followed more than 20,000 adolescent substance users over time as they received access to and treatment with promising evidence-based treatments Dennis et al. (2003) such as the Adolescent Community Reinforcement Approach (A-CRA) Godley (2001) and Motivational Enhancement Treatment/Cognitive Behavior Therapy, 5 Sessions (MET/CBT5) Sampl and Kadden (2001). Specifically, we will illustrate how to estimate the relative effectiveness of these two treatment programs for adolescents on total days abstinent 6-months later.

Estimation of accurate causal treatment effects Holland (1986) is the main goal of many observational studies including our motivating case study. The causal effect of a treatment for each individual is defined as the difference in the outcome for an individual had they received that treatment compared to the outcome had they not received it. This is practically impossible to measure directly since typically, only one treatment condition will be assigned to each individual Rosenbaum and Rubin (1983). As a consequence, we end up with longitudinal, observational study data, where two (or more) groups receive different treatments over time, and then we estimate the causal treatment effect from the difference in the outcomes of the groups. Without adjustments for confounding, however, this estimate will almost certainly be biased unless we control in a meaningful way for the pretreatment group differences between the groups.

The propensity score (PS) Holland (1986) is the probability of an individual's allocation to a specific treatment group, given their observed baseline (pre-treatment) characteristics. The PS can be used to create comparable treatment groups when we have observational study data by either weighting, matching, adjusting or stratifying on the PS. By minimising the imbalance of known and observed confounders between the treatment groups, PS

methods reduce the bias in the estimation of the causal treatment effect due to the observed confounders. There are numerous approaches to using the PS to draw causal inference. Each approach has its merits and its challenges. Matching Ho et al. (2007), King, G., and Nielsen, R. (2019), Stuart (2010), involves using the PS to match individuals of one group to individuals of the other group, while stratification Leite (2016) involves creating bins of individuals with similar levels of the PS (e.g., $0-0.2, 0.2$–0.4, etc.) so that individuals within each stratum are more similar to each other. In contrast, PS weighting creates pseudo-populations with similar baseline characteristics, and control for over-sampling - individuals with similar characteristics within one group Leite (2016). Finally, there is also the option to adjust for the PS as a control covariate Myers and Louis (2010). Here we focus on the use of PS weighting to estimate causal effects Hernán et al. (2000), Robins et al. (2000) (as opposed to PS matching, stratification and adjustment) given the notable increase in methods for estimating weights that have arisen Elze et al. (2017), Harder et al. (2010), Olmos and Govindasamy (2015), Posner and Ash (2012).

In addition to PS weights, we also consider the closely related balancing weights developed via entropy balancing Hainmueller (2012). Entropy balancing computes the balancing weight directly, as opposed to traditional PS algorithms, which first compute the PS and then transform it to balancing weights. In practice, it is impossible to know aprior which of these candidate methods will be able to balance treatment groups best for a given case study. Thus, we have developed a tutorial which includes all of the methods jointly by estimating the needed balancing weights using all the methods and then selecting the one that is optimal for a given case study by comparing the balance achieved and the precision that will be achieved by the methods (as captured via the effective sample size).

As noted, there are several estimation methods for both PS and balancing weights, including parametric and non-parametric modeling and machine learning techniques. However, there is no clear indication that a single method performs best in every dataset Griffin et al. (2017), Setodji et al. (2017), Setoguchi et al. (2008). This is the reason, we suggest, that one should consider multiple methods and finally estimate the causal treatment effect based on the one that best balances the treatment groups without reducing the effective sample size unduly. This tutorial will present an implementation of some of the more commonly used PS estimation methods, namely Logistic Regression (LR) Agresti (2018), Wright (1995), Generalized Boosted Model (GBM) McCaffrey et al. (2004), and Covariate Balance Propensity Score (CBPS) Imai and Ratkovic (2014). We will also consider the use of balancing weights via Entropy Balancing (EB) Hainmueller (2012), Zhao and Percival (2016) and provide a summary of the advantages and disadvantages of each approach. Our goal is not to compare the performance of these algorithms Zhao and Percival (2016), rather to highlight the importance of using all of them as potential tools to compute balancing weights. The performance of the methods will be illustrated by applying them to understanding the relative effects of A-CRA versus MET/CBT5 on days abstinent for adolescent substance users.

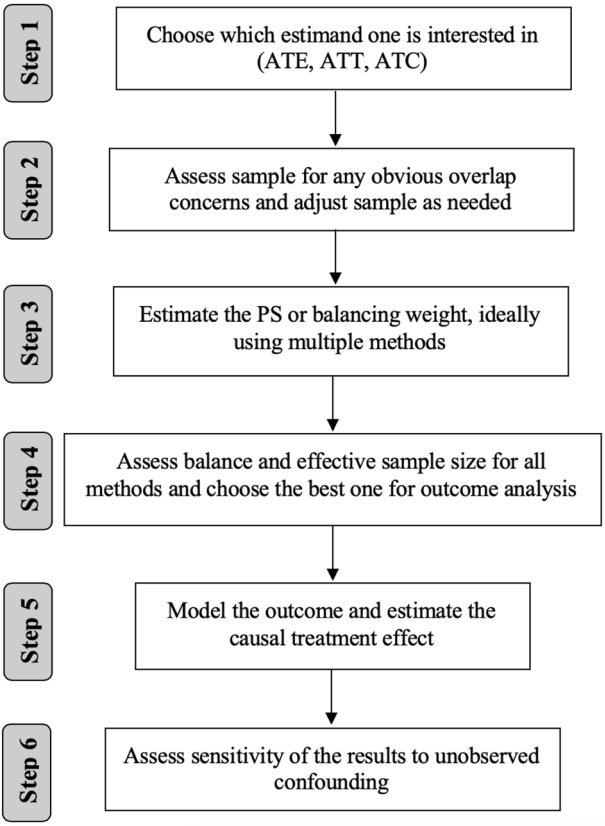
Graphical tools have been used in the past to assess the balance between the baseline covariates of the treatment groups Garrido et al. (2014), Stuart (2010), and these graphs could also be used here to assess the groups for lack of common support and/or potential outliers. In this tutorial, we attempt to implement these steps by using density plots to examine the support of the covariates between the treatment groups, before the application of the balancing weights. We then adjust the sample if needed, without concern about the balance between the groups at this point. On a later step, balance statistics will be used to examine the balance achieved between the baseline covariates of the treatment groups, using the Propensity Score and balancing Weights (PSBW).

Our tutorial will also carefully consider two key assumptions Ho et al. (2007), Stuart (2010) required for estimation of causal effects using PS or balancing weights, namely that the two groups have sufficient overlap in the observed pretreatment covariates and there are no unobserved confounders missing from the analysis. Both the overlap and unobserved confounding assumptions are impossible to test formally. The consideration of these assumption have been previously discussed Ho et al. (2007), Stuart (2010), however, here we present useful steps one can take in practice to assess the potential validity of the overlap and the potential impact unobserved confounders might play in a given analysis. As such, we make these two steps critical parts of our six steps. Overall, our tutorial offers the field an updated and more complete sets of steps for estimation of high-quality PS and balancing weights when one is working with binary treatments. Our use of multiple estimation methods and careful consideration of key assumptions is unique from prior tutorials in the field Bergstra et al. (2019), Lee and Little (2017), McCaffrey et al. (2013), Olmos and Govindasamy (2015), Ridgeway et al. (2017). Here, we highly recommend the consideration of several methods and to make inference based on the one that achieves the best trade-off between balance and effective sample size (ESS). Finally, the examination of the sensitivity of the estimation of the causal treatment effect to unobserved confounder — violation of one of the key assumptions of causal modelling — assesses the impact of model design on power of the findings.

The remainder of the article is organised as follows. First, we summarizes the six key steps needed to estimate causal treatment effects of a treatment in observational studies. Then, we explain the PS and EB estimation algorithms we use in detail as well as the multiple measures of performance we use to identify the optimal methods. Finally, we apply the six steps to our motivating case study comparing the relative effectiveness of two evidence-based treatment programs for substance use for adolescents. We note that we have developed a user-friendly Shiny application that can be used to run all steps proposed in this tutorial on any data set. It can be found at https://andreasmarkoulidakis.shinyapps.io/cobweb/.

2 The 6 Key Steps towards Estimating Causal Treatment Effects

There is a wide discussion in the literature, regarding the number of steps necessary to estimate causal treatment effects using balancing and PS weights Bergstra et al. (2019), Caliendo and Kopeinig (2008), Setodji et al. (2017). Here, we follow 6 key steps, uniquely considering the relative performance of several estimation methods for the balancing and PS weights as well as demonstrating the needed and, often underutilized, use of omitted variable sensitivity analyses.



Step 1 *Choose which estimand one is interested in (ATE, ATT, ATC).*

In order to define the commonly used causal treatment effects, we use the potential outcomes notation first introduced by Rubin (1974). For each individual in a study with two groups, we define Y_{1i} to denote their potential outcome under treatment and Y_{0i} to denote their potential outcome under control for $i = 1, ..., N$. While these potential outcomes exist in theory for all individuals in our study, we only get to observe one potential outcome for each individual. Namely, we observe the outcome for the treatment they actually ended up receiving. We define our treatment indicator as T_i where values of 1 denote that individual i received treatment and 0 denotes receipt of the control condition. Then we can define $Y_i^{obs} = Y_{1i} \cdot T_i + Y_{0i} \cdot (1 - T_i)$.

The most commonly used causal treatment effects (following notation of Leite (2016)) are:

- The *Average Treatment Effect on the Entire population (ATE)*

$$ATE = E[Y_{1i}] - E[Y_{0i}], \tag{1}$$

- The *Average Treatment Effect on the Treated population (ATT)*

$$ATT = E[Y_{1i} \mid T_i = 1] - E[Y_{0i} \mid T_i = 1], \tag{2}$$

- The *Average Treatment Effect on the Control population (ATC)*

$$ATC = E[Y_{1i} \mid T_i = 0] - E[Y_{0i} \mid T_i = 0]. \tag{3}$$

Each estimand is used for a different purpose, depending on the research question a study is trying to answer. ATE allows one to understand the average causal treatment effect for the entire population of individuals in both the treatment and control groups. In contrast, ATT allows one to understand the effect of the treatment among only individuals like those in the treatment group, and ATC to understand the effect of the treatment among only individuals like those in the control group. For example, considering our case study, the ATE would measure the relative effect of the two treatments for all adolescents in our sample, while ATT and ATC would quantify the relative effect of the two treatments for adolescents like those in the individual treatment groups. If we assume A-CRA is officially to be defined as the "treatment" group and MET/CBT5 as the "control/comparison" condition, then ATT measures the relative effectiveness of the two programs for youth like those in the A-CRA group while ATC measures the relative effectiveness of the programs for youth like those in the MET/CBT5 group. Since adolescents in the two groups tend to be different in several ways (e.g., youth in the A-CRA group were older, more likely to be a minority and had higher mean levels of substance use and mental health problems and past substance use and mental health treatment than youth in the MET/CBT5 group), the populations to which ATT and ATC generalize will be different.

Step 2 *Assess sample for any obvious overlap concerns and adjust as needed.*

Rubin's Causal Model (RCM) is the first widely known approach to the statistical analysis of causal treatment effects, considering potential (not necessarily observed) outcomes. It is based on two critical assumptions Rosenbaum and Rubin (1983) — *stable unit treatment value assumption* (SUTVA) and *strong ignorability*. SUTVA implies that the potential outcomes for each individual are independent of the treatment status of the other individuals Cox and Cox (1958). Strong ignorability includes two key parts related to the PS (or treatment assignment mechanism), defined here as $Pr(T_i = 1| \mid |\mathbf{X}_i)$ where $\mathbf{X}_i$ denotes the vector of observed pretreatment confounders used in the PS model. First, that there are no unobserved confounders in the PS model. We discuss how to address this assumption in more detail in Step 6. The second part states that each individual has a positive probability to be assigned to the treatment group ($0 < Pr(T_i = 1| \mid |\mathbf{X}_i) < 1$) Rosenbaum and Rubin (1983). In observational studies, where the group assignment is likely to be determined by the medical and personal characteristics of the individual, it is possible to be able to predict the group allocation perfectly based on the baseline characteristics Bergstra et al. (2019) if certain groups of individuals in the study only ended up in one group versus another. For this reason, it is important to report some summary statistics of the baseline covariates (like mean, standard deviation, minimum, maximum) per group to check that the distributions of the covariate values of the groups overlap. Unfortunately, there is no formal way to test this overlap assumption in a given sample. Instead, we recommend some simple checks that can be done to assess obvious areas of the covariate distributions where there is a lack of overlap. For example, overlap can be checked by comparing the minimum and maximum of the same covariate in the two groups or by distribution plots such as those

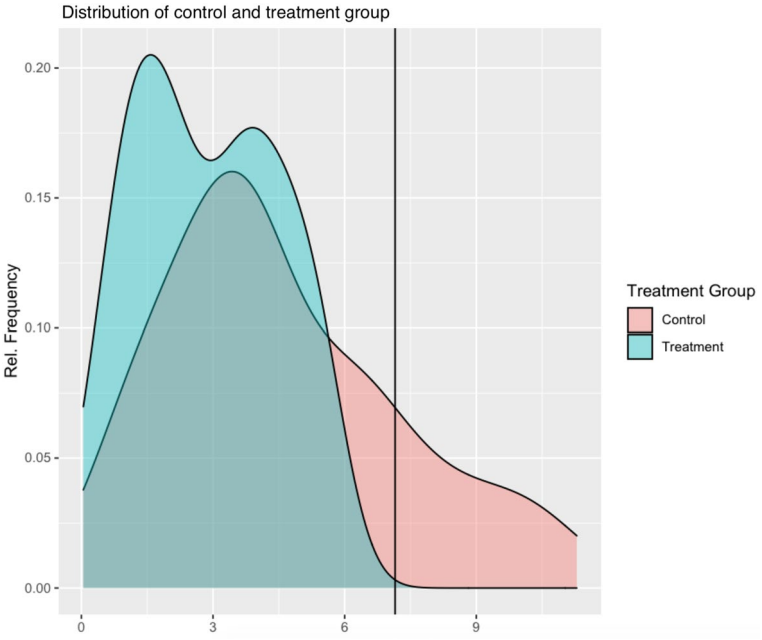


Fig. 1 Density plot of control and treatment group, where there is lack of overlap in the support of the two groups. The data used for this plot are artificial to demonstrate the issue

shown in Fig. 1. If there are ranges of covariate values that only occur in one group, one might have to exclude some individuals from the analysis, such that the two groups are adequately overlapped, or consider other estimands, like the *Average Treatment Effect on Overlapping Population* (ATO) Li et al. (2018), Mlcoch et al. (2019), which measures the causal treatment effect on the region where there are representatives of both groups. Lack of overlap suggests that there is no uncertainty in the decision to treat individuals with a particular characteristic since individuals in this region are always assigned to one treatment group. The exclusion of participants with covariate values that do not overlap with the other group tends to create a more clinically useful target population. In the case of binary or/and categorical covariates, one should make sure that at least one individual is representing each category/level per treatment group — unless a category has no representative in either group[1].

We use Fig. 1 to illustrate how plotting the density functions of a single covariate for each of the treatment groups can be a helpful way to identify obvious areas of the covariate distributions lacking overlap. It is apparent from Fig. 1 that the support of the control group

[1] When there is a significant lack of overlap on the support of the treatment groups for the baseline covariates, it is not feasible to estimate the ATE with low bias Li et al. (2018). ATO provides an asymptotically unbiased alternative, with the minimum asymptotic variance. As the ATO weights are considered only for regions where the supports overlap, the balancing weights are bounded and can achieve close balance (as measured by the difference of the means of the treatment groups), even when PS is estimated using logistic regression and the sample size is small Li et al. (2018). Even though ATO seems to have desirable properties, interpretation is specific to each application, and sometimes this means lack of generalization of the outcomes to key target populations for studies or trials.

(red area) extends beyond the support of the treatment group — this is the area to the right of the vertical line. In such a case, one should consider how to handle this issue, either by estimating the balancing weight only for the common area — in practice, this could mean to exclude control group individuals with large values for this covariate — or by estimating the treatment effect only for the region of overlap (ATO) Li et al. (2018).

Consideration of overlap prior to estimation of the balancing and PS weights is an important part of the process of estimating causal treatment effects. Often researchers dive into estimation of the balancing and PS weights without careful consideration of overlap and find themselves frustrated by obtaining balancing weights that do not successfully balance their groups. If lack of overlap exists, it can be difficult to obtain high quality balancing and PS weights. Our Shiny application allows the user to examine the summary statistics of every confounder, as well as their density plots. There is an option to trim outliers for either upper or lower tail, or both, for one confounder at a time or more. This procedure enables the user to view the impact of removal of observations from one confounder on the remaining confounders when the observation is removed from the entire data set. It is quite common that observations with extremes in one confounder tend to be similarly extreme for other confounders.

The use of density plots aims to observe any obvious outliers, or areas of no support overlapping (no overlap on the x-axis) of the treatment groups on the baseline covariates, rather than modelling these densities exactly (particularly in the case of the categorical age covariate, which is rounded to the nearest year). To assess the a-priori balance of the treatment groups, we highly recommend assessing the balance among the baseline covariates prior to the weighting. In certain cases, the baseline characteristics are adequately balanced, thus no adjustment is needed. In such a scenario, using no balancing weights is beneficial, in the sense that any weights could reduce the ESS, while no weights would maintain the power of using the original data unrefined.

Variable Selection Discussion about which variables will be included in the PS model to control for confounding bias is beyond the scope of this tutorial. It is recommended in the literature to prioritize the true confounders — this is covariates related both the treatment allocation and the outcome —, and then, if the sample size allows, to include covariates related only to the outcome, as this could reduce the Mean Square Error (MSE) of the causal treatment effect estimand Markoulidakis et al. (2021).

Step 3 *Estimate the propensity score or balancing weights needed, ideally using multiple methods*

There have been several articles comparing PS and balancing weighting methods (for example, Abdia et al. (2017), Griffin et al. (2017), Mao et al. (2019), Setodji et al. (2017), Setoguchi et al. (2008)). Under different settings, different methods perform better, and this depends on the structure of the given data set — the sample size, the number of covariates to be balanced (especially relative to sample size), and the true underlying form of the treatment assignment model (e.g., linear versus non-linear). If one has a small sample and the natural relation between the confounders and the allocation mechanism is a sigmoid function with main effects only, logistic regression seems a natural choice. In general, however, there is a lack of guidance on how to choose from the multitude of methods available in any particular analysis, in a way to achieve sufficient covariate balance without reducing the effective sample size more than necessary, to achieve (near) unbiased estimation of a causal treatment effect. It is not immediately obvious in any given setting which method is optimal. Thus, in this tutorial, we recommend the consideration of several methods and to

make inference based on the one that achieves the best trade-off between balance and effective sample size (ESS) as explained in more detail in Step 4 Ridgeway et al. (2017).

Assessment of PS Distribution Overlap There are several references that recommend checking the distribution of PS across the two groups D'Agostino Jr (1998), Elze et al. (2017), McCaffrey et al. (2004), Li et al. (2018) since lack of overlap could lead to extreme weights and potentially compromised estimation of the causal treatment effect. A region where the distributions of PS of the two groups do not overlap means that there is no uncertainty about the treatment allocation of individuals in this region — this is that the second assumption of the RCM does not hold. Regions of lack of overlap on PS distributions are regions in which the baseline characteristics of these individuals predict exactly the treatment allocation of an individual. This is why it is important to assess the overlap of the baseline covariates in step 2, to avoid such problematic behaviors in the PS distributions. Given we address this issue in step 2, we consider that checking the PS distributions is not always necessary, as there should likely be no lack of overlap.

Step 4 *Assess balance and effective sample size for all methods and choose the best one for outcome analysis*

Once PS or balancing weights are estimated, the balance (or comparability) among the groups needs to be assessed. The theory behind PS and balancing weights suggests that balance should be obtained on the full multivariate distribution of the observed confounders after one applies the weights. However, in practice, this is often not checked and it is challenging to properly test if the full multivariate distribution of the observed confounders balances between the treatment conditions.

Here, we propose to use both the *standardized mean difference* (SMD) and the *Kolmogorov-Smirnov statistic* (KS) as a way to assess how comparable the two treatment groups are. The SMD allows one to assess the comparability of the means for each observed confounder while the KS statistic allows us to assess balance also in the tails of the distributions for a given confounder between the two treatment groups. These metrics are commonly used in the literature Austin (2009), Franklin et al. (2014), Gail and Green (1976), Griffin et al. (2017), Setodji et al. (2017), Setoguchi et al. (2008), Zhang et al. (2019). Both metrics are explained in detail in Sect. 4.

Additionally, we also carefully consider the impact of the weighting on the power of a study. The PS and balancing weights act in the same way as survey or sampling weights and add increased variability into the statistical models and treatment effect estimates. We can assess the impact of different PS and balancing weight methods by computing the ESS, which denotes the remaining sample size, after the reduction due to the variability in the weights. When balance across multiple methods is similar, one would naturally prefer to select as optimal the PS or balancing weights with the lowest reduction to the ESS. The ESS is also explained in more detail in Sect. 4.

When applying PSBW to control for confounding bias on the estimation of causal treatment effects, we perform balance checked after weighting. For reference, the app accompanying this tutorial, reports the balance values of the treatment groups prior to the weighting, but this is only for reference.

Step 5 *Model the outcome and estimate the causal treatment effect.*

Before proceeding with any outcome analyses, researchers must assess whether they will be able to estimate robust causal effects with their sample. To do so, they need to ensure adequate balance has been obtained with the PS or balancing weights being used in a given analysis. If the weights do not balance the groups being compared well[2], the estimation of the causal treatment effect could be highly biased Markoulidakis et al. (2021). It is important that researchers understand when this is not happening, the study can only be used to examine associations, rather than causal effects, since the findings will be less robust and must be caveated as such. Assuming one does have adequate balance and the largest possible ESS, there are several possible options for the estimation of the needed treatment effects. The simplest estimate is to compare weighted means between the treatment and control groups. Given the balancing weights w_i, for every individual i, an estimation of the causal treatment effect could be obtained by the formula:

$$\frac{\sum_{i \in C_1} w_i Y_i^{obs}}{\sum_{i \in C_1} w_i} - \frac{\sum_{i \in C_0} w_i Y_i^{obs}}{\sum_{i \in C_0} w_i} \tag{4}$$

where C_1 is the set of individuals in the treatment group, and C_0 the set of individuals in the control group.

However, it is more common practice to combine the weights with a multivariable regression adjustment that ideally includes all of the observed confounders used in the estimation of the weights Austin (2011), Ridgeway et al. (2017). That is, the PS or balancing weights are used as weights in an augmented regression model that also includes all observed confounders. For PS weights, this approach yields a *doubly robust estimator* of the causal treatment effect Bang and Robins (2005), Kang et al. (2007), Zhao and Percival (2016). The estimated treatment effect is consistent so long as one part of the doubly robust model is correct (i.e., either the PS weight model or the multivariable outcome model). Our Shiny application uses this approach to estimate the causal treatment effect of interest. For balancing weights, this doubly robustness property has not yet been established. However, the use of covariate adjustment in the regression model is still seen as useful for minimizing bias in the estimated treatment effect and increasing precision in the model. In cases where sample sizes are restricted and might not support fully adjusting for all covariates, it can be useful to control for a subset of the observed confounders, namely those that have the greatest lingering imbalance in the SMD or KS statistic.

Step 6 *Assess sensitivity of the results to unobserved confounding.*

A key assumption in all weighted analyses is that we have not left out any potential unobserved confounders when estimating the PS or balancing weights. Unfortunately, as with the overlap assumption, the assumption of no unobserved confounders is impossible to test formally in practice. Yet, it is important to assess how robust a study's findings might be to unmeasured factors that have not been included in the weights. It is most common to utilize sensitivity analyses that assess the sensitivity of the estimated treatment effects and/or statistical significance of analysis to potential unobserved confounders. Despite their importance, such analyses are underutilized in the literature. Here, we showcase the use

[2] Two, or more, groups are considered to be sufficiently well balanced as soon as the values of the SMD and KS statisticare both below 0.1.

of a graphical tool to describe how sensitive both treatment effect estimates and statistical significance (as measured by the p-value) will be to an unobserved covariate Griffin et al. (2020).

The method works by using simulations to assess how both the estimated treatment effect (solid contours) and p-value (dashed contours) would change as a function of an unobserved confounder whose association with the treatment indicator is expressed through an effect size or SMD (displayed on the x-axis), and whose relationship with the outcome is expressed as a correlation (displayed on the y-axis). For each point on the graph (e.g., for a fixed assumed SMD relationship between the unobserved confounder and treatment and a fixed assumed correlation relationship between the unobserved confounder and the outcome), the method generates an unobserved confounder for each individual in the sample that meets these assumed strengths of relationships and then updates the balancing weights so that the weight for each individual now properly reflects inclusion of the unobserved confounder. This update takes the form of multiplying the balancing weight by a factor that is a function of the unobserved confounder and the assumed SMD relationship between the unobserved confounder and treatment. These updated weights are used to re-estimate the treatment effect and associated p-value in an effort to understand how much the two change as we increase the strength of the relationships between the unobserved confounder and both treatment and the outcome (Burgette et al; Forthcoming on arXiv). Naturally, all studies will be sensitive to an unobserved confounder and the method used here conveniently helps put the findings into perspective by plotting the observed pre-treatment confounders onto the graph to showcase the type of observed strengths of relationship that already exist in the dataset. If the results of a study are such that they are highly likely to change within levels of the observed confounders used in the balancing weights, one must be concerned that the findings from the study might be highly sensitive to an unobserved confounder. In contrast, if the results are relatively stable within the region of the observed pre-treatment covariates, then one can feel that the findings are likely robust to the presence of an unobserved confounder. The method used is highly related to the e-value VanderWeele, et al. (2017), though utilizes a graphical representation of how the findings might change over a range of different assumed relationships between the unobserved confounder and treatment and the outcome to tell a fuller story.

3 Propensity scoring and balancing weight analysis methods

We will now introduce some notation, which will be useful in the description of PS and balancing weight estimation algorithms. Consider a simple random sample of N observations from a population P. For each unit i, we observe a binary treatment variable T_i and a vector of observed pre-treatment covariates $\mathbf{X}_i$. The PS is defined as the conditional probability of receiving the treatment given the covariates $\mathbf{X}_i$, i.e.$Pr(\mathbf{X}_i) = Pr(T_i = 1| \, | \, |\mathbf{X}_i)$. From Rosenbaum and Rubin (1983), the ignorability of treatment assignment says that the treatment assignment is ignorable given the (true) propensity score $Pr(\mathbf{X}_i)$. This implies that the unbiased estimation of treatment effect is possible by conditioning on the PS alone instead of the entire covariate vector $\mathbf{X}_i$. However, in observational studies the PS must be estimated from the data set. Normally, one assumes a parametric PS model $Pr_\beta(\mathbf{X}_i)$. That is, $Pr(T_i = 1| \, | \, |\mathbf{X}_i) = Pr_\beta(\mathbf{X}_i)$, where β is a vector of unknown parameters.

In the following paragraphs, four commonly used algorithms for obtaining PS and balancing weights are described.

1. **Logistic Regression**

The simplest and most commonly used parametric model method to estimate the PS of each individual is logistic regression (LR) Agresti (2018) since treatment assignments are often binary. The basic logistic regression model for estimating the PS assumes that the *logit* of the probability of receiving treatment is equated with a linear combination of covariates $X_i^T \beta$. The coefficient vector β is usually estimated with maximum likelihood estimation. The PS, in this case, is then computed from the estimated parameters as follows:

$$Pr_{\beta}(X_i) = \frac{exp(X_i^T \beta)}{1 + exp(X_i^T \beta)} \tag{5}$$

An extension of LR is multinomial logistic regression which could be used to estimate generalized PS if there are more than two treatment conditions. The main problem with the LR approach is that the PS model can easily be mis-specified, leading to biased estimates of treatment effects Lee and Little (2017), Pirracchio et al. (2015), Setoguchi et al. (2008), Wyss et al. (2014). It is nearly impossible to know the best way to specify the right hand side of the LR model.

2. **Covariate Balancing Propensity Score**

To overcome the shortcoming caused by mis-specification of the model and create a parametric option focused on achieving good balance between the treatment groups, the covariate balancing PS method was developed by Imai and Ratkovic (2014). The authors used the covariate balancing property of parametric models by employing inverse PS weighting:

$$E\left[\frac{T_i f(X_i)}{Pr_{\beta}(X_i)} - \frac{(1 - T_i) f(X_i)}{1 - Pr_{\beta}(X_i)}\right] = 0 \tag{6}$$

where $f(X_i)$ is a measurable function of X_i specified by the researcher. For instance, if $f(X_i)$ is the first derivative of $\pi_{\beta}(X_i)$, the assumed parametric model is logistic. The above equation holds for the estimation of the ATE, while modified versions deployed for ATT and/or ATC Imai and Ratkovic (2014).

This method has the advantage of being robust to mild model mis-specification with regard to balancing confounders compared to direct maximum likelihood estimation used in a standard LR. Additionally, the CBPS method can improve the covariate balance in observed data sets and improve the accuracy of estimated treatment effects over parametric models even if there is no mis-specification Choi et al. (2019), Imai and Ratkovic (2014), Setodji et al. (2017), Wyss et al. (2014), Xie et al. (2019).

The covariate balance method uses a generalized method of moments or an empirical likelihood estimation approach to find estimates that come closest to optimizing the likelihood function while concurrently meeting the balance condition for the weighted means of the covariates in the parameter estimation procedure. Even though the default set of restrictions (expressed through $f(\cdot)$) targets to balance the first moment (this is to minimize

the SMD Austin and Stuart (2015) — see *section Measures to Evaluate Balance*), one could apply higher moments restrictions. Second-order restrictions are easily implemented, by transforming each continuous confounder via an orthogonal polynomial of degree 2 Huang and Vegetabile (2017). This procedure increases the number of baseline covariates included in the treatment allocation model (times m, where m is the order of the orthogonal polynomial), thus special caution should be taken here since this could make the achievement of adequate balance harder. *CBPS*#1, *CBPS*#2 and *CBPS*#3, denote the use of CBPS matching on the first, second and third moments of the distributions of the baseline covariates. Respectively, these correspond to matching the mean, variance and skewness of the distribution. Finally, in the application in this article, over-identified CBPS was deployed, which means that the restrictions imposed on the confounders, were greater in number than the number of the parameters of the model Imai and Ratkovic (2014)[3]

3. **Generalized Boosted Model**

GBM is a flexible, nonparametric machine learning approach to estimating PS weights. It predicts the binary treatment indicator by fitting a piecewise-constant model, constructed as a combination of simple regression trees (Burgette, McCaffrey, and Griffin in press, Burgette et al. (2015), Ridgeway (1999)), namely *Recursive Partitioning Algorithms* and *Boosting*. To develop the PS model, GBM uses an iterative, *forward stagewise additive algorithm*. Starting with the PS equal to the average of treatment assignment on the sample, such an algorithm starts by fitting a simple regression tree to the data to predict treatment from the covariates by maximizing the following function.

$$l(x) = \sum_{i=1}^{N} T_i g(X_i) - log(1 + exp(g(X_i))), \tag{7}$$

where $g(X_i)$ is the *logit* of treatment assignment. Then, at each additional step of the algorithm, a new simple regression tree is added to the model from the previous iterations without changing any of the previous regression tree fits. The new tree is chosen to provide the best fit to the residuals of the model from the previous iteration. This chosen tree also provides the greatest increase to the log-likelihood for the data. When combining trees, the predictions from each tree are shrunken by a scalar less than one to improve the smoothness of the resulting piecewise-constant model and the overall fit.

The number of iterations that are performed by the algorithm or the number of trees in the model determines the model's complexity. When choosing the number of iterations to yield the final PS model, one must pick a value that balances between underfitting (i.e. not capturing important features of the data) and overfitting the data. One selects the *final* model of the treatment indicator (and correspondingly, the PS and PS weights needed for analysis) by selecting a particular number of iterations considered *optimal* where optimization is done based on achieving the best balance — the most commonly used editions of the algorithms are the ones who target to optimize the mean standardized mean difference (GMB_{ES}) and the maximum Kolmogorov-Smirnov statistic (GBM_{KS}). A detailed tutorial of GBM for PS estimation, balance evaluation, and treatment effect estimation in R software is discussed in Ridgeway et al. (2017). GBM methods could also be used when there are more than two treatments McCaffrey et al. (2013).

[3] This estimation method of CBPS offers better asymptotic efficiency, but could perform poorly on restricted samples, where the use of just-identified is recommended by Imai and Ratkovic (2014).

4. **Entropy Balancing**

(EB) is a method that aims to estimate the weights directly rather than via the PS of the individuals. The method promises to achieve exact balance on as many moments (e.g., the mean, variance, skewness) as defined by the user. Assuming that one is interested in estimating the ATT, the quantity that is tricky to estimate is $E[Y_{0i}| \,|T_i = 1]$ (this is the expected outcome value of individuals in the control groups, considered they received the treatment), since these values are not observed. Thus, an estimation of the above quantity is

$$\hat{E}[Y_{0i} \mid T = 1] = \frac{\sum_{i=1}^{n_0} w_i^0 Y_i^{obs}}{\sum_{i=1}^{n_0} w_i^0}, \tag{8}$$

where w_i^0 are the balance weights **for the individuals in the control group** and need to be estimated, and n_0 is the number of individuals in the control group. The entropy balancing method calculates weights through a re-weighting scheme (until adequate balance in the pre-advised moments is achieved), while at the same time attempting to satisfy a set of *balance and normalizing constraints* — Eqs. (3), (4) and (5) in Hainmueller (2012). This set of restrictions corresponds to matching the first k moments of the distributions of the two groups (k is defined by the user), which results in matching more characteristics of the distributions of the baseline covariates between the treatment groups — here, we compute PSBW using $k = 1, 2$ and 3, as with CBPS algorithm — denoting $EB\#1, EB\#2$ and $EB\#3$, respectively,to denote when the EB algorithm matches on the first moment (the means), the first and second moments (consequently matching the mean and variance) and the first three moments (consequently matching the mean, variance, and skewness), respectively.

Entropy balancing re-weighting schemes can be considered as generalizations of the traditionally used Inverse Probability Weighting (IPW) Hainmueller (2012), Zhao and Percival (2016), Zhao et al. (2019). The algorithms described above, estimate the PS of an individual to be assigned to the treatment group, conditional to their baseline characteristics. IPW is the mechanism traditionally used to convert the PS (probabilities of assignment to the treatment group) to balancing weights (see Eqs. (9), (10), and (11)) — weights which balance the baseline characteristics of the treatment and control group. These PSs are computed, in each case, considering a set of restrictions (depending on the algorithm). EB, belongs to a class of algorithms, that computes a set of weights directly considering a set of restrictions concerning the balancing achieved between the treatment group, rather than the PS of each observation — thus a direct estimation of PS is not feasible in this case.

The algorithms discussed above (namely LR, CBPS, GBM and EB), are computing PS and balancing weights on a different manner. LR, CBPS and GBM are estimating the PS of allocation to the treatment group for each individual based on their baseline characteristics — the first two fitting parametric models, while the later via regression trees. LR sets no restrictions, while the CBPS and GBM do use a set of restrictions concerning the balance of the treatment and control groups — this depends on the form of $f(\cdot)$ for CBPS, and the stopping method for GBM. For instance, when the ES_{mean} is used as a stopping method for GBM, the algorithm is estimating the PSs, trying to minimize the mean value of SMD between the treatment groups. This is why, CBPS and GBM require to specify whether one is interested in ATE, ATT or ATC, so that the PS are converted to balancing weights. EB, is estimating the balancing weights directly, based on a set of restrictions concerning the number of moments (k) the user wishes to match — if $k = 1$ the algorithm will compute weights attempting to balance the means of the baseline characteristics of the treatment groups, if $k = 2$ the weights will be estimated in such a way that the means values and the

means of the square values of the baseline characteristics are equal for the two groups (thus the mean and variances of the two groups), etc..

We note that the PS weights w_i are defined in a different way, depending on the components of Eqs. (1), (2) or (3) we wish to estimate. The weights are defined as follows:

$$w_i = T_i \frac{1}{Pr(x_i)} + (1 - T_i)\frac{1}{1 - Pr(x_i)}, \quad for \quad ATE, \tag{9}$$

$$w_i = T_i + (1 - T_i)\frac{Pr(x_i)}{1 - Pr(x_i)}, \quad for \quad ATT, \tag{10}$$

$$w_i = T_i \frac{1 - Pr(x_i)}{Pr(x_i)} + (1 - T_i), \quad for \quad ATC. \tag{11}$$

Logistic regression, *GBM* and *CBPS* produce PS, which we transform into balancing weights using Eqs. (9), (10) and (11), whereas *entropy balancing* estimates the weights directly.

All the key-steps required for the estimation of causal treatment effect, were implemented using the CoBWeb app Markoulidakis et al. (2021). For the computation of the PSBW, the app uses 1. *CBPS* Ratkovic (2013) for CBPS algorithm, controlling for the first $m = 1, 2, 3$ moments; 2. *twang* Ridgeway et al. (2017) for GBM algorithm; and 3. *entbal* Vegetabile Brian (2021), for entropy balancing, controlling for the first $m = 1, 2, 3$ moments. *OVtool*Pane et al. (2021) is used to produce the sensitivity analysis plot.

4 Measures to evaluate balance

Once the estimation of weights is available, it is important to evaluate the relative performance of the methods. To do so we will rely on three key metrics: *standardized mean difference* (SMD), *Kolmogorov-Smirnov statistic* (KS) and *effective sample size* (ESS). The first two are measures of the balance of observed covariates, while the latter is a measure of the sample power lost by weighting.

Standardized Mean Difference SMD Austin (2009), Austin and Stuart (2015), Franklin et al. (2014), Li et al. (2018) is a measure of the distance of the means of two groups. It is defined as the difference of the means, divided by an estimate of the standard deviation, for a given covariate — depending on the causal treatment effect one wishes to estimate. For ATE, it is formally defined as:

$$SMD = \frac{\bar{X}^{weighted}_{treatment} - \bar{X}^{weighted}_{control}}{\sqrt{\frac{sd^2_{treatment}}{n_0} + \frac{sd^2_{control}}{n_1}}}, \tag{12}$$

where $X^{weighted}_{treatment}$ and $X^{weighted}_{control}$ are the weighted sample means. A weighted mean is defined as $\bar{x}^{weighted} = \frac{\sum_i w_i x_i}{\sum_i w_i}$, where w_i is the weight of observation i — in our context this is the PS and balancing weights. As the variance within each treatment group impacts the SMD, in the denominator of the above equation, we recommend the use of $sd_{treatment}$ and $sd_{control}$, which are the unweighted standard deviations (SD) of treatment and control group, as opposed to the weighted SD of the two groups, since weights usually inflates variances. n_0

and n_1 are the sample size of the treatment and control group, respectively. Alternatively, one could use weighted s.d., however, inflated variance could diminish SMD Austin (2011). In this article, we report the absolute value of the SMD (absolute standardized mean difference) Li et al. (2018).

The lower the SMD value, the better the balance achieved. In the recent literature, 0.1 is recommended as a threshold Austin (2008), Austin (2009), Austin et al. (2007), Austin and Mamdani (2006), Griffin et al. (2017), Griffin et al. (2014), Ho et al. (2007), Stuart et al. (2013), Zhang et al. (2019) to define groups as balanced, while in the past 0.2 was considered as adequate balance. Abdia et al. (2017), however, provides strong evidence that such a threshold is very liberal, and does not guarantee an unbiased estimator of the causal treatment effect. By definition, SMD quantifies the similarity of the means of the two groups for each variable, thus it expresses how close the two mean values are.

Kolmogorov-Smirnov Statistic KS is a statistical test Gail and Green (1976) which checks the hypothesis that the two samples are from the same distribution. Its test statistic is formally defined as

$$KS = \max_{z} |F_{emp}^{treatment}(z) - F_{emp}^{control}(z)|, \tag{13}$$

where $F_{emp}^{treatment}(\cdot)$ and $F_{emp}^{control}(\cdot)$ are the empirical distributions of treatment and control, respectively. The empirical distribution of a sample $x_1, x_2, ..., x_n$, is:

$$F_{emp}(x) = \frac{\#x_i \leq x}{n}. \tag{14}$$

By definition, the *KS statistic* takes values in [0, 1], and the lower the value, the closer the two distributions. Unlike SMD, it is a measure that quantifies the similarity of the entire distribution of the two groups, rather than the means only. It is apparent from (14), that the value of KS statistic (Eq. (13)) relies partially on the sample size of each treatment group, thus adequate balance (low KS values) are typically easier to achieve when a large sample is available Markoulidakis et al. (2021).

The value of SMD is not bounded, as it measures the distance of the mean values, compared to the standard deviation of the treatment groups. In contrast, the KS statistic takes values in [0, 1]. Typically, 0.1 is used as a threshold for SMD Austin (2008), Austin (2009), Austin et al. (2007), Austin and Mamdani (2006), Griffin et al. (2017), Griffin et al. (2014), Ho et al. (2007), Stuart et al. (2013), Zhang et al. (2019),. There is no clear guidance on the optimal threshold for KS, thus it is not as commonly used as a measure of balance of the treatment groups. There is evidence that reducing the KS statistic below 0.1 reduces the bias in treatment effect estimates Markoulidakis et al. (2021). Therefore, we propose to use 0.1 as the threshold for balance for the KS as well as the SMD.

The definitions in Eqs. (13) and (14) are not considering weights, but this could be easily incorporated, if we replace the sample $\{x_i\}_i$ with $\{z_i\}_i$, which is the weighted version of $\{x_i\}_i$ — $z_i = \frac{w_i}{\sum_i w_i} x_i$, for every i.

Effective Sample Size Given the weights of each group, the ESS Ridgeway et al. (2017), Shook-Sa and Hudgens (2022) is defined as:

$$ESS = \frac{\left(\sum_{i \in C} w_i\right)^2}{\sum_{i \in C} w_i^2}, \tag{15}$$

where w_i are the weights of the group C — this could be either treatment or control group, if we are interested in the estimation of ATC or ATT, respectively, or the entire sample, if we wish to estimate ATE.

ESS expresses the number of observations from a random sample one would have to use to obtain an estimate with the same variance as the one obtained from the weighted group. It, therefore, can be used to help understand the power/precision a study has after using PS or balancing weights. The ESS will always be smaller than the *original* sample size.

Extensive comparisons of the algorithms discussed in Sect. 3, in terms of balance achieved, is beyond the purpose of this article. However, bias in the estimation of the causal treatment effect is minimized if adequate balance has been achieved, whichever method is used Markoulidakis et al. (2021). Thus, it is important to be cautious when we pick which balancing weights will be used for the estimation of the causal treatment effect, choosing the ones computed by the algorithm that achieves the best balance. Different algorithms impose different balance restrictions to estimate PSBW (LR imposes no restrictions), and this often has an impact on the ESS. Using multiple algorithms provides the ability to consult the ESS when adequate balance is achieved by multiple algorithms and select as optimal the method that maximizes ESS.

5 Study data

As noted, the motivating case study data used in this tutorial comes from longitudinal observational study data on adolescents receiving substance use disorder (SUD) treatment, who were administered the Global Appraisal of Individual Needs (GAIN) biopsychosocial assessment instrument Dennis et al. (2003) on a recurring basis. The GAIN was routinely collected by 178 adolescent SUD treatment sites funded by the Center for Substance Abuse Treatment (CSAT) between 1997 and 2012, including the sites serving our A-CRA and MET/CBT5 groups. In total, we have 4968 adolescents from the A-CRA group and 5184 from the MET/CBT5 group with GAIN data at baseline and 6-months post-baseline.

Since the allocation mechanism between the two treatment groups is not random, we will control for 20 pretreatment confounders (for explanation of these covariates see Dennis et al. (2003)) in our analysis. In brief, they cover a range of baseline confounders including sociodemographic factors like age, race/ethnicity, and gender as well as baseline measures of substance use, mental health and environmental risk. These confounders were chosen based on prior research with experts in the field of SUD Grant et al. (2020) as well as the literature as key potential predictors of outcomes in this sample Griffin et al. (2012), Ramchand et al. (2015), Ramchand et al. (2014), Ramchand et al. (2011), Schuler et al. (2014). While dealing with variable selection for PS and outcome model is beyond the goal of this current tutorial Brookhart et al. (2006), Hirano and Imbens (2001), we note that it is important to ensure selection of all important potential confounders prior to performing the analyses outlined in this tutorial. Our key outcome is the total number of days abstinent in the 90-days prior to the 6-month follow-up.

Table 1 shows the baseline characteristics of youth in our two groups. As shown, the groups showed differences on 17 out of our 20 pretreatment confounders (i.e. they have SMD and/or KS > 0.1). In particular, Table 1 suggests A-CRA more likely to be non-white. They also had higher mean levels of substance use and mental health problems (including traumatic stress, emotional problems, internal mental distress and behavioral complexity). Youth in

Table 1 Mean values of baseline characteristics per group

	MET/CBT5	A-CRA	Balance	
	Mean	Mean	SMD	KS
Age	15.42	15.62	0.15	0.07
Traumatic Stress Scale	1.79	2.26	0.14	0.06
Substance Frequency Scale	10.40	11.88	0.12	0.05
Depressive Symptom Scale	2.59	2.80	0.08	0.04
Emotional Problems Scale	20.73	25.48	0.25	0.12
Internal Mental Distress Scale	7.49	8.51	0.12	0.06
Behavior Complexity Scale	9.62	11.02	0.17	0.07
Adjusted Days Abstinent (past 90 days)	55.11	47.76	0.22	0.10
In recovery	0.25	0.25	0.00	0.00
Mental Health Treatment (past 90 days)	0.18	0.23	0.11	0.05
Substance Abuse TX Index	1.44	8.01	0.39	0.14
General Conflict Tactic Scale	2.85	3.23	0.14	0.07
Continued Substance Use Despite Prior Tx	0.09	0.11	0.07	0.10
Recovery Environmental Risk Index	47.85	41.94	0.35	0.16
Parent Activity Index	3.57	3.26	0.23	0.11
Substance Use Dependence (past year)	3.64	4.39	0.22	0.10
Living Environmental Risk Index	10.32	10.78	0.15	0.06
Social Environmental Risk Index	12.80	13.22	0.10	0.05
	% per group		Balance	
	MET/CBT5 (%)	A-CRA (%)	SMD	KS
Race:White	50.8	31.9	0.41	0.19
Race:Hispanic	22.5	31.8	0.21	0.09
Race:African American	10.1	15.9	0.17	0.06
Race:other	16.6	20.5	0.10	0.04
Gender:Female	26.75	29.9	0.07	0.03

A-CRA were also more likely to have had past substance use and mental health treatment than youth in the MET/CBT5 group. Finally, youth in the A-CRA group were less likely to have parental involvement at home and more likely to have higher levels of risk in their living and social environments. In light of these differences between the treatment groups on the baseline pretreatment confounders, it is clear this case study will benefit from use of PS or balancing weights to remove the differences between the two groups on these observed baseline factors. We will now walk through the six needed steps of this tutorial to obtain balancing weights that ensure the groups are more comparable and showcase how to robustly examine the estimated treatment effects from the optimal set of weights.

We note that the original case study data analysis included 90 imputed datasets. We randomly selected one of the 90 imputed data sets to allow for easier illustration of the steps of this tutorial. When doing PS or balance weighting with imputed data, it is necessary to repeat steps 4 and 5 of the tutorial for each imputed data set and then aggregate results across imputations to get the final estimated effect of treatment.

6 Obtaining PS weighting estimates and assessing their performance

The Covariate Balancing & Weighting Web App (*CoBWeb*) was developed by the main author, as part of the article, to help users apply the procedure described in *section The 6 Key Steps towards Estimating Causal Treatment Effects*. The app implements all the steps, in a user-friendly way. *CoBWeb* is freely available at https://andreasmarkoulidakis.shinyapps.io/cobweb/.

General guidance about the implementation of the *key steps* in the app is described in the Appendix.

6.1 Step 1: Choose which estimand one is interested in (ATE, ATT, ATC)

This analysis will focus on estimating the ATE in order to allow us to generalize the findings for our sample of youth enrolled in both treatments.

6.2 Step 2: Assess sample for any obvious overlap concerns and adjust sample as needed

Next, we carefully assess the overlap between the control (MET/CBT5) and treatment (A-CRA) groups. Notably, there were no concerns about overlap for any of the 20 pretreatment confounders used in our analysis. Figure 2 provides a view of the overlap density plots for three of the confounders – age, substance use at baseline, and the depression scale – to illustrate the patterns seen for all 20 confounders. As shown, there are no concerns about lack of overlap for these confounders. Both groups have individuals with a similar range of observed values for each confounder. As it is apparent, the treatment groups are well overlapped, thus there is no apparent need to trim the data set. We will continue the analysis without the removal of any outliers. Finally, from the last two columns of Table 1, it is apparent that the treatment groups are not balanced at the baseline — Emotional Problems Scale, Substance Abuse TX Index, Parent Activity Index, and White Race, report KS-statistic values above the threshold of 0.1.

6.3 Step 3: Estimation of propensity scores or balancing weights, ideally using multiple methods

Now we estimate the balancing and PS weights using the four methods presented in section *Propensity Scoring and Balancing Weight Analysis Methods* (this corresponds to 9 algorithms in total). Here, we are interested in the estimation of ATE, thus the relevant formula for estimation of needed PS weights from LR, GBM and CBPS is given by Eq. (9). EB estimates the balancing weights directly and, thus, no transformation is done.

6.4 Step 4: Assess balance and effective sample size for all methods and choose the best one for outcome analysis

Having estimated the balancing weights, we next assess balance after applying the weights from each method.

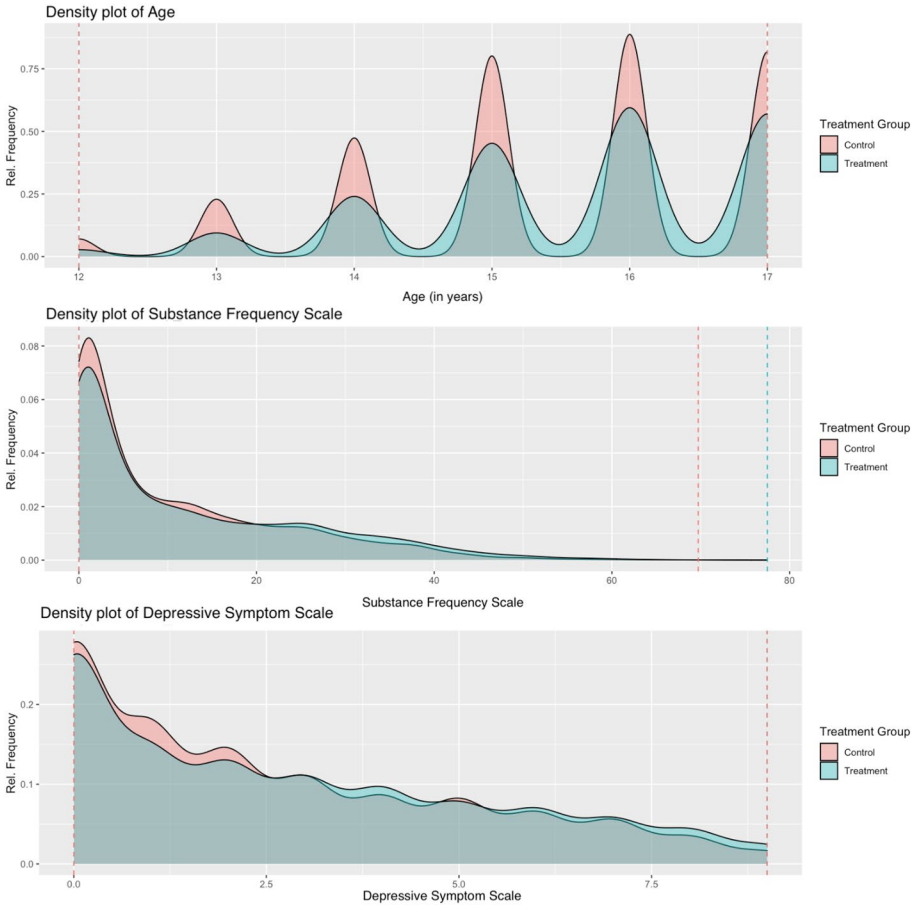


Fig. 2 Density plots of *Age*, *Substance Frequency Scale* and *Depressive Symptom Scale* — light red is the control group (A-CRA) and with light blue the treatment group (MET/CBT5)

Table 2 reports the mean and the maximum SMD and KS values per method, as well as the ESS per method.

As observed in Table 2, *EB* achieves the best balance in terms of SMD (absolutely 0 for all covariates — since maximum SMD is equal to 0), followed closely by all other algorithms. Only the logistic regression model has a maximum SMD (0.13) above the widely accepted threshold of 0.1 Ridgeway et al. (2017), which suggests there is lingering meaningful imbalance for the PS weights associated with this method.

In terms of KS statistic, the lowest values are reported by *GBM* algorithms — maximum KS value equal to 0.03 for both algorithms. All algorithms though, achieve adequate balance when using the 0.1 threshold for balance.

Taken all together, the majority of our approaches achieved successful balance after weighting except for logistic regression. All have maximum SMD and KS below the 0.1 cut-point — but LR.

At this point, it would be advisable to choose the optimal algorithm as the one with the best combination of lowest maximum KS and the largest ESS — the sample size retained

Table 2 Standardized Mean Difference (SMD), Kolmogorov-Smirnov Statistic (KS) and Effective Sample Size (ESS) per estimation method.

	SMD		KS		ESS	% of original Sample
	Mean	Max	Mean	Max		
Unweighted	0.15	0.39	0.07	0.16	10152	100%
LR	0.02	0.13	0.02	0.05	6062	60%
GBM_{ES}	0.03	0.07	0.02	0.03	7752	76%
GBM_{KS}	0.03	0.07	0.02	0.03	7877	78%
CBPS#1	0.01	0.02	0.02	0.06	7534	74%
CBPS#2	0.01	0.05	0.01	0.05	7564	75%
CBPS#3	0.02	0.06	0.01	0.04	7711	76%
EB#1	0	0	0.01	0.05	8071	80%
EB#2	0	0	0.01	0.04	7838	77%
EB#3	0	0	0.01	0.03	7743	76%

For SMD and KS 0.1 is considered as a widely accepted threshold for good balance, while ESS should be as close to unweighted as possible. *CBPS*#1, *CBPS*#2 and *CBPS*#3 refer to estimation of balancing weights using the CBPS algorithm, controlling for the first (*CBPS*#1), first and second (*CBPS*#2), first, second and third (*CBPS*#3) moments, respectively. *EB*#1, *EB*#2 and *EB*#3 refer to estimation of balancing weights using the EB algorithm, controlling for the first (*EB*#1), first and second (*EB*#2), first, second and third (*EB*#3) moments, respectively

after the weighting process. Although, with the exception of LR, differences between the algorithms are small, GBM_{KS} appears optimal by this criterion. Thus, balancing weights estimated by this algorithm will be used for the outcome analysis.

6.5 Step 5: Model outcome and estimate the causal treatment effect

Once adequate balance is achieved, the next step is to estimate the causal treatment effect, using the balancing weights identified as optimal in the previous step.

To do so, we fit a weighted linear regression on *Adjusted Days Abstinent (past 90 days)* as the outcome, measured at 6-months after intake, considering as predictors all the other pretreatment covariates shown in Table 1 (including *Adjusted Days Abstinent (past 90 days)* measured at baseline), and the treatment status (A-CRA). The coefficient of treatment status represents the estimand of interest. The treatment effect is estimated to be equal to 1.3, with a 95% confidence interval equal to (0.038, 2.562) The estimation of the CI here is based on quantiles from the standard normal distribution, due to the large size of the sample. and an associated p-value of 0.043 ($\simeq 0.05$), which means that the effect of the treatment is considered to be small and marginally statistically significant (at 5% level of significance), suggesting that youth in A-CRA had slightly higher mean days abstinent at 6-months post intake (here approximately equal to 1.3 days more with a 95% confidence interval equal to (038, 2.562) days.

The estimation of the standard error of the causal treatment estimator is done based on the formulas used to estimate the standard error of regression coefficients, as the actual treatment effect is estimated via regression models Bang and Robins (2005), Kang et al. (2007), Zhao and Percival (2016).

6.6 Step 6: Assess sensitivity of the results to unobserved confounding

In the final step of the tutorial, we carefully consider how sensitive findings from the case study might be to potential omitted pretreatment confounders. Specifically, we use a graphical tool Pane et al. (2021) which describes how sensitive both treatment effect estimates and statistical significance (as measured by the *p*-value) will be to an unobserved covariate.

The output is a contour plot (Figure 3), which shows how both the estimated treatment effect (solid contours) and *p*-value (dashed contours) would change as a function of an unobserved confounder whose association with the treatment indicator is expressed through an effect size or SMD (x-axis) and whose relationship with the outcome is expressed as a correlation (y-axis). More specifically, the x-axis displays the *Association with Treatment Indicator* — this is the unweighted standardized mean difference of the potential unobserved confounder between the treatment and control groups —, and the y-axis displays the *Absolute Association with the Outcome Covariate* — this is the absolute correlation of the unobserved confounder with the outcome covariate. The plot also provides the estimated treatment effect and the associated *p*-value (caption) from the original analysis, which is an indicator of the significance of the effect.

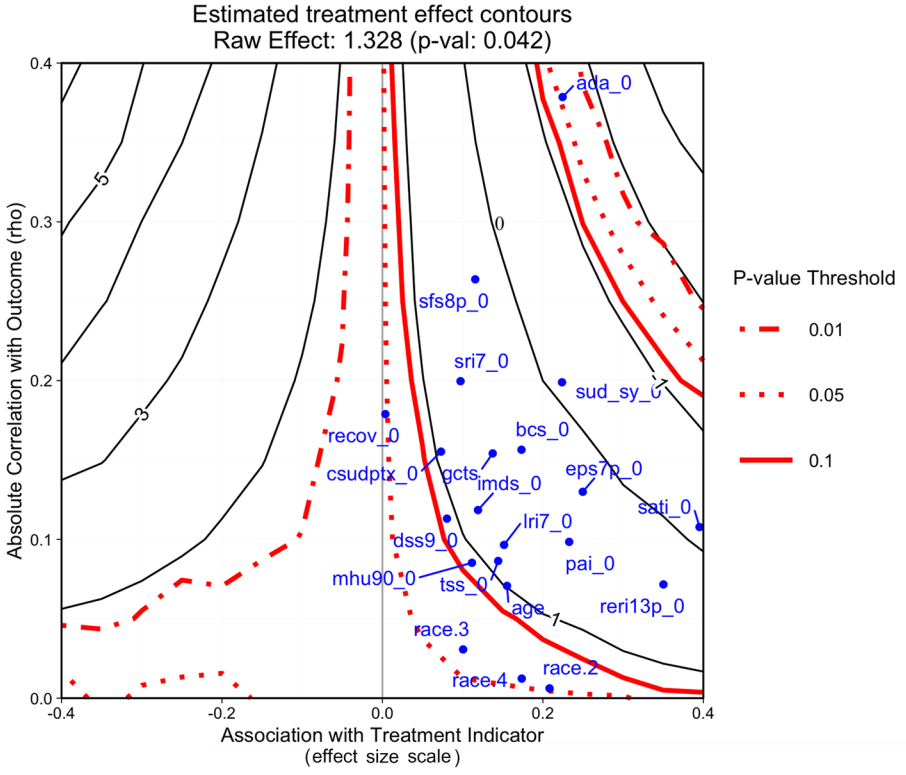


Fig. 3 The x-axis indicates the SMD of potential confounders, and the y-axis the correlation with the outcome. The *black solid* contours represent the size of the treatment effect, while the red dashed ones represent the *p*-value cut-offs (0.01, 0.05, 0.1 levels of significance). The blue dots represent the association of the confounders with the treatment and the outcome

The solid black lines of the contour plot show how the size of the treatment effect changes as a function of the unobserved confounder's relationship with treatment assignment and the outcome, here showing that the estimated treatment effect will get larger as we move to the left from 0 on the x-axis but will decrease to 0 and then move negative as we move right from 0 on the x-axis.

The red contours show how the size of the *p*-value changes as a function of the unobserved confounder focusing on commonly used cut-offs. Since our finding had a *p*-value near the 0.05 threshold, we find out that the statistical significance only gets stronger as we move left from 0 on the x-axis and the estimated treatment effect gets larger. As we move right from 0 on the x-axis, we quickly move into an area of statistical non-significance **(the area between the solid red lines)** before again crossing back into an area where the *p*-value will be less than 0.05 as our treatment effect starts getting larger in the negative direction.

The blue dots help to put the findings into perspective by plotting the observed correlations between our observed pretreatment confounders and the outcome and treatment indicator. Here, these all fall into the range where our treatment effect is close to 0 and has a non-significant *p*-value. This suggests it is highly possible we left out an important unobserved confounder that might quickly make our estimated treatment effect basically null. An example unobserved confounder that might have correlations similar to the observed confounders shown would be family history of substance use or mental health problems.

Our findings in this case study are clearly *very sensitive* to unobserved confounders. As a consequence, considering an unobserved confounder would be negatively associated to the treatment (negative SMD on the baseline) on the range [0, 0.4] — possessed on the left side of the graph—, and a low/moderate association to the outcome (correlation with the outcome in the range [0, 0.4]), would influence the estimated causal treatment effect towards 0 value, with lower significance, and thus provide evidence that the days abstinent under A-CRA would not be considered significantly different compared to MET/CBT5.

7 Discussion/Conclusion

In this paper, we present a step-by-step guide to making inference on causal treatment effects using observational data. This tutorial represents an advancement over prior work Ali et al. (2016), Garrido et al. (2014), Lee and Little (2017), Olmos and Govindasamy (2015) in that we explicitly deal with addressing two key assumptions for PS methods (overlap and unobserved confounding) and recommend the use of multiple estimators for the potential PS and balancing weights in order to ultimately select the best method for a given study. It is difficult to be able to project which PS or balancing method will do best in any given study and thus, we believe it is better for it to become standard practice (before looking at any outcome models) to use multiple methods and carefully compare balance and ESS to select the one that is optimal.

Observational studies are widely used in the research of causal treatment effects. Thus, the accurate estimation of the effect of interest is of primal importance. Since the treatment and control groups are not balanced a priori, unlike in a (large) RCT, PS and balancing weights are a useful tool for every researcher who wishes to make inference based on observational data.

Observational studies, though, can pose several challenges, including extreme values on baseline covariates in one group, as well as limitations related to sample size. When dealing with small samples, a case which is quite often when studying rare diseases, one should be careful with the number of confounders which one attempts to balance, because there is always the danger of model overfitting. Variable selection is beyond the scope of this tutorial. We used expert input and past literature to choose the covariates we are interested to control for in our case study, however, there is some literature covering the topic of variable selection for PS models Brookhart et al. (2006), Hirano and Imbens (2001).

It is crucial to investigate the data *a priori* for potential overlapping or outliers issues, and adjust as needed. We highly recommend considering more than one method when it comes to algorithms that produce balancing weights and then evaluate the balance for each method and select the best performing one. Following outcome analysis it is important to perform an analysis of the sensitivity of the outcome estimation and significance to unobserved confounding, to access the generalization abilities of the outcome results.

A CoBWeb

CoBWeb Markoulidakis et al. (2021) is freely available at https://andreasmarkoulidakis.shinyapps.io/cobweb/

Initially, the user has to upload a data set using the *Data* panel. A summary of each covariate will appear (mean, standard deviation, median, minimum and maximum value), as well as the first few rows of the raw data.

A.1 Choose which estimand one is interested in (ATE, ATT, ATC)

Moving to the *Model Set Up* panel, there will appear some multiple-choice questions to define the *Treatment Status* covariate, any other binary, continuous and categorical confounders, the outcome covariate, — all the covariates included in the dataset will be available options — as well as a choice of the estimand that the user wishes to estimate (ATE, ATT, ATC).

A.2 Assess sample for any obvious overlap concerns and adjust sample as needed

The *Dealing with Outliers* panel provides a view of the per-group summary statistics of each covariate — mean, sd, median, min, and max for three of the confounders. Under the sub-panel, the user will have a glance at the differences of the distributions of the two treatment groups, for each covariate, which provides a graphic way to access any obvious overlapping concerns — options for removing outliers are available here.

A.3 Estimation of propensity scores or balancing weights, ideally using multiple methods

Next, we estimate balancing weights using the four methods presented in Section *Propensity Score and Balancing Weight Analysis Methods* (9 algorithms in total).This is automatically done in the app.

A.4 Assess balance and effective sample size for all methods and choose the best one for outcome analysis

In the next panel of the app, called *Balance Evaluation*, the SMD, KS statistic, and ESS (raw and as a percentage of the original size) are reported for each PS and balancing weights algorithm, alongside some recommendations about which algorithm performs better in each measure. At this stage, the user can choose which algorithm they wish to proceed with for the outcome analysis.

A.5 Model outcome and estimate the causal treatment effect

The *Outcome Analysis* panel is split into two sub-panels — *Treatment Effect Estimation* and *Sensitivity Analysis*. The former provides a table with the estimation of the causal treatment estimation effect and its statistical significance.

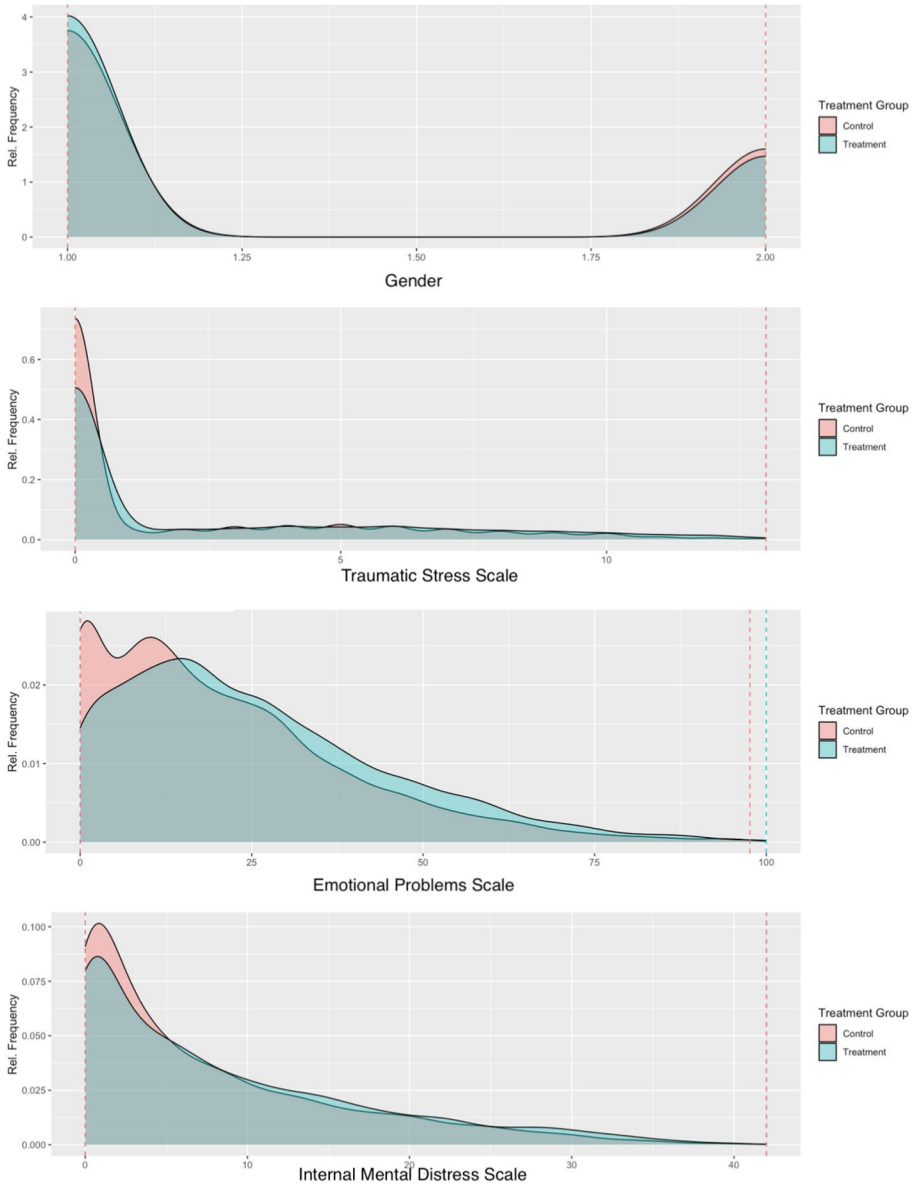
A.6 Assess sensitivity of the results to unobserved confounding

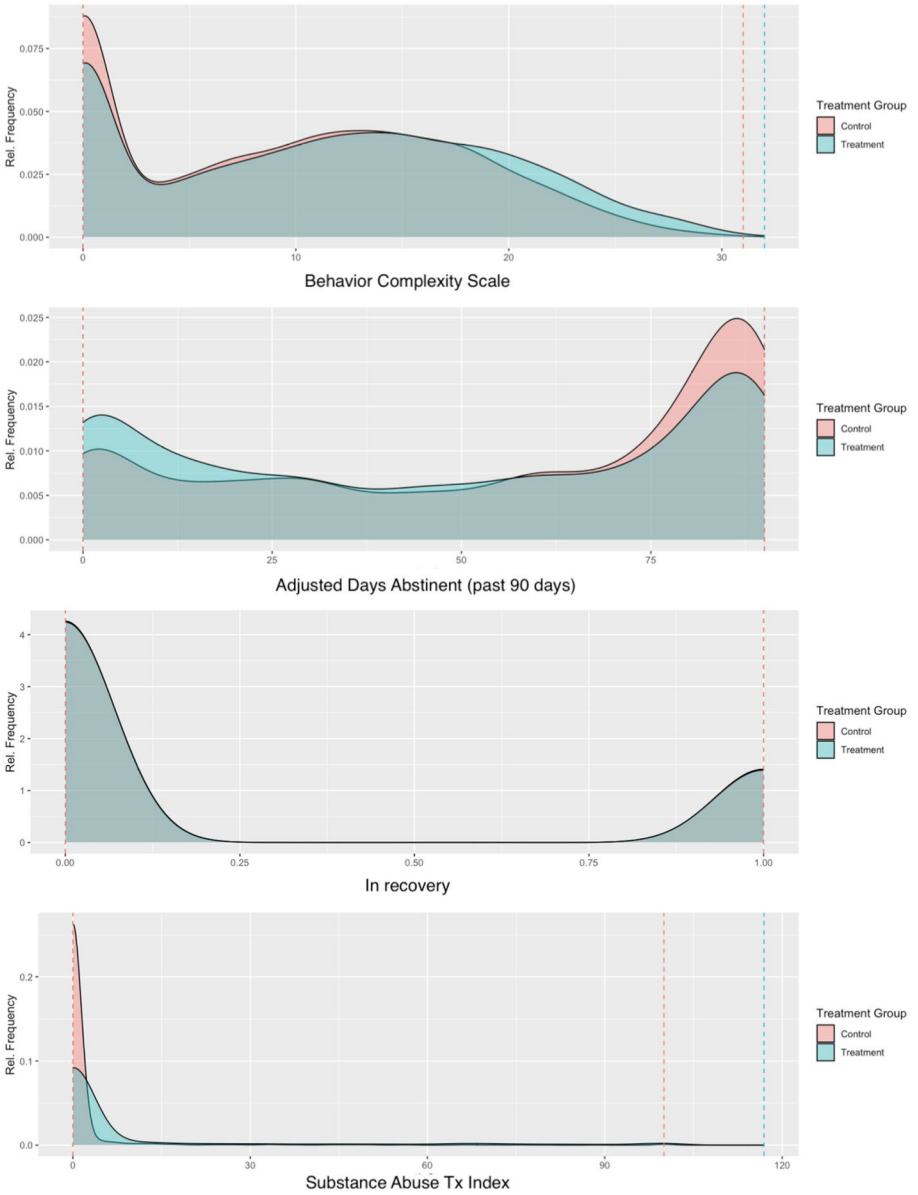
If one wishes to proceed with the sensitivity analysis (rather than stopping after step 5) — which is something we highly recommend — this can be done in the *Sensitivity Analysis* sub-panel of the *Outcome Analysis* panel. Once this is done, the user can download the sensitivity analysis graph, and move to the final panel (*Before you go!*).

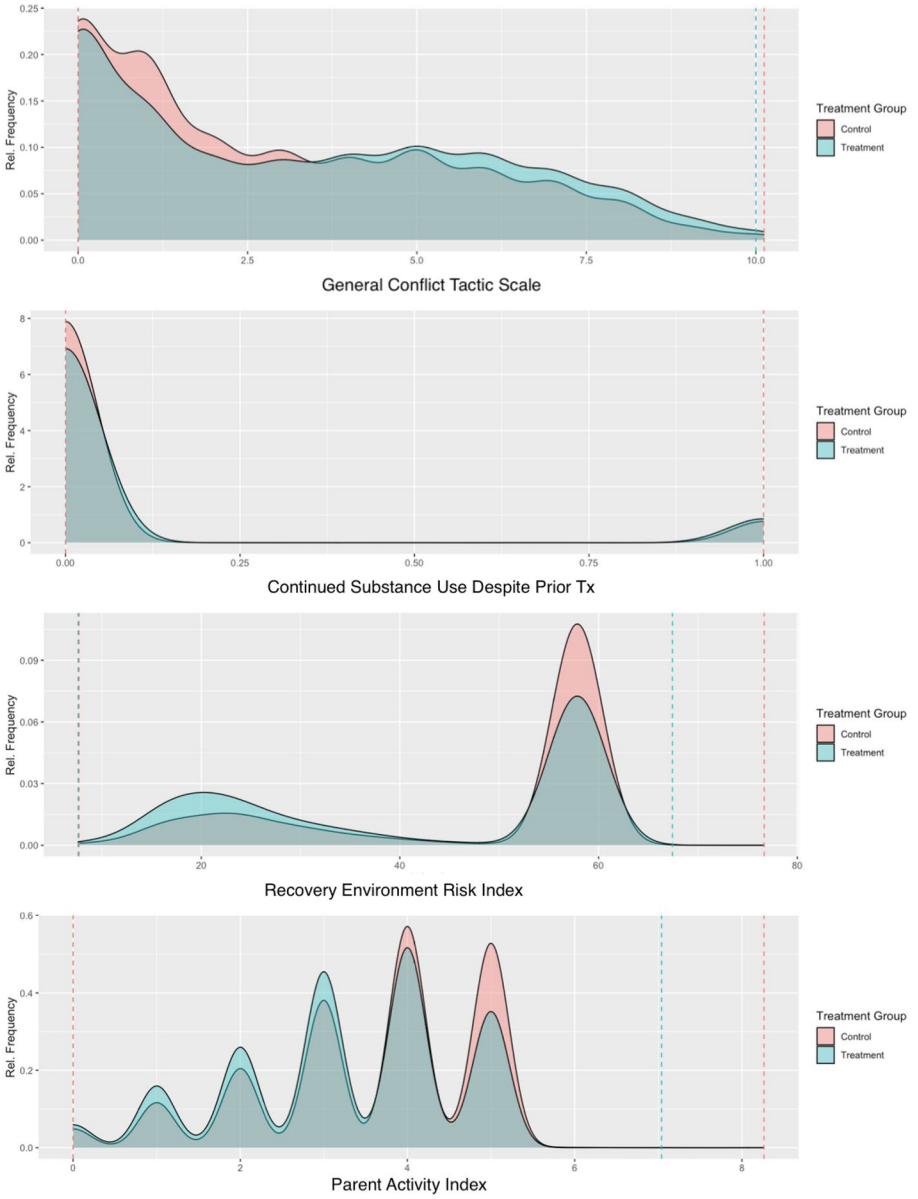
Before you go!

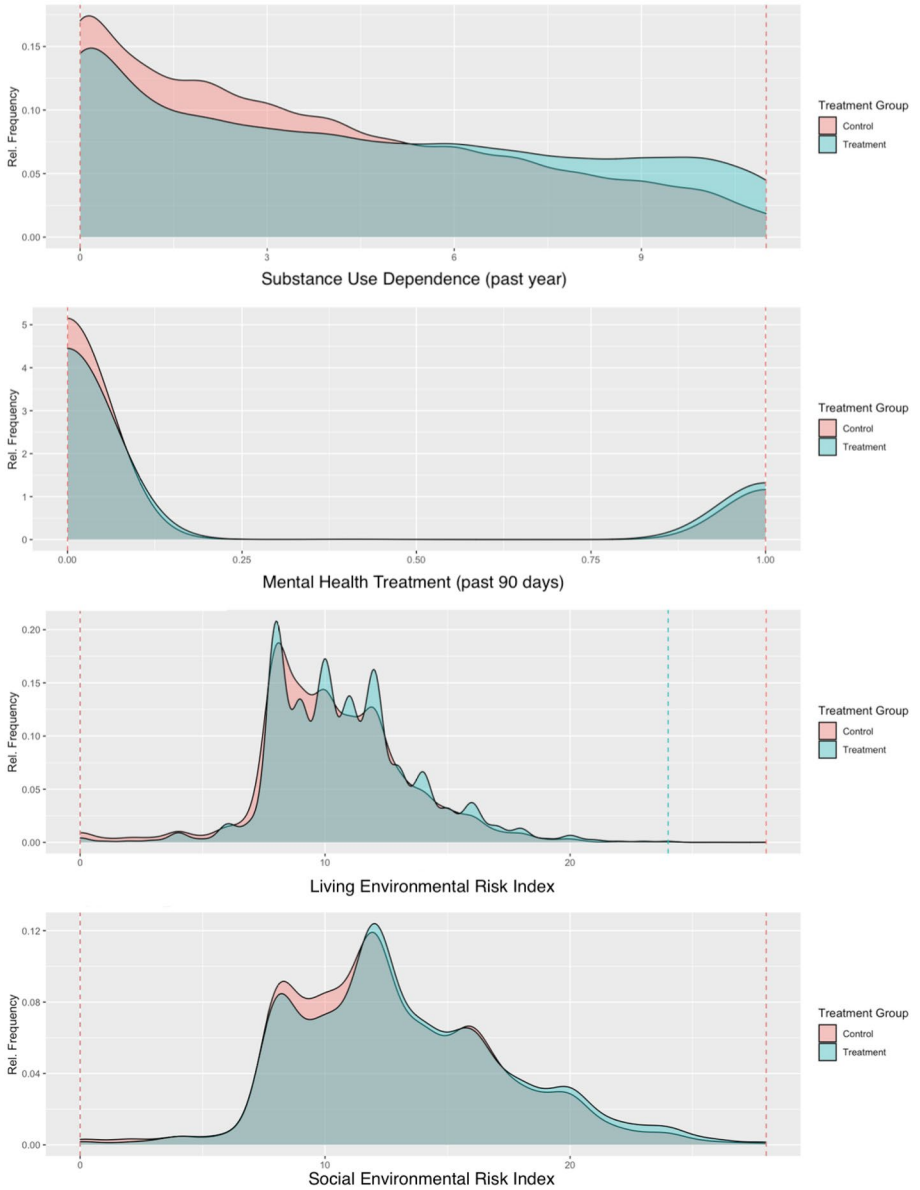
In the final panel (*Before you go!*) one can download an enhanced report (including the sensitivity analysis findings), and the final data with the weights.

B Density plots




Rel. Frequency
0.075
0.050
0.025
0.000
0
10
20
30
Behavior Complexity Scale
Treatment Group
Control
Treatment
Rel. Frequency
0.025
0.020
0.015
0.010
0.005
0.000
0
25
50
75
Adjusted Days Abstinent (past 90 days)
Treatment Group
Control
Treatment
Rel. Frequency
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
In recovery
Treatment Group
Control
Treatment
Rel. Frequency
0.2
0.1
0.0
0
30
60
90
120
Substance Abuse Tx Index
Treatment Group
Control
Treatment


Rel. Frequency
0.25
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
General Conflict Tactic Scale
Treatment Group
Control
Treatment
Rel. Frequency
8
6
4
2
0
0.00
0.25
0.50
0.75
1.00
Continued Substance Use Despite Prior Tx
Treatment Group
Control
Treatment
Rel. Frequency
0.09
0.06
0.03
0.00
20
40
60
80
Recovery Environment Risk Index
Treatment Group
Control
Treatment
Rel. Frequency
0.6
0.4
0.2
0.0
0
2
4
6
8
Parent Activity Index
Treatment Group
Control
Treatment



Acknowledgements The development of this manuscript was also supported by the Center for Substance Abuse Treatment (CSAT), Substance Abuse and Mental Health Services Administration (SAMHA) [#270-2003-00006, #270-2003-00006, and #270-2007-00004C] using data provided by the following grantees: TI-15413, TI-15415, TI-15421, TI-15433, TI-15438, TI-15446, TI-15447, TI-15458, TI-15461, TI-15466, TI-15467, TI-15469, TI-15475, TI-15478, TI-15479, TI-15481, TI-15483 TI-15485, TI-15486, TI-15489, TI-15511, TI-15514, TI-15524, TI-15527, TI-15545, TI-15562 TI-15577 TI-15584, TI-15586, TI-15670, TI-15671, TI-15672, TI-15674, TI-15677, TI-15678, TI-15682, TI-15686, TI-17589, TI-17604, TI-17605, TI-17638, TI-17646, TI-17648, TI-17673, TI-17702, TI-17719, TI-17728, TI-17742, TI-17744, TI-17751, TI-17755, TI-17761, TI-17763, TI-17765, TI-17769, TI-17775, TI-17779, TI-17786, TI-17788, TI-17812,

TI-17817, TI-17821, TI-17825, TI-17830, TI-17831, TI-17847, TI-17864, TI-20759, TI-20781, TI-20798, TI-20806, TI-20827, TI-20828, TI-20847, TI-20852, TI-20865, TI-20870, TI-20910, TI-20946, TI-23174, TI-23186, TI-23188, TI- 23195, TI-23196, TI-23197, TI-23200, TI-23202, TI-23204, TI-23206, TI-23224, TI-23244, TI-23247, TI-23265, TI-23270, TI-23276, TI-23278, TI-23279, TI-23296, TI-23298, TI-23304, TI-23310, TI-23311, TI-23312, TI-23316, TI-23322 ,TI-23323 ,TI-23325 ,TI-23336 ,TI-23345 ,TI-23346, TI-23348. The authors thank these agencies, grantees, and their participants for agreeing to share their data to support this secondary analysis. The opinions about this data are those of the authors and do not reflect official positions of the government or individual agencies. Please direct correspondence about the data to Beth Ann Griffin, 1200 South Hayes Street, Arlington, VA 22202, USA bethg@rand.org, (703) 413-1100x5188.

Funding DOMINO-HD (which funds the Ph.D. of the main author) is funded through the EU Joint Program for Neurodegenerative Disease Research with UK funding from Alzheimer's Society and Jacques and Gloria Gossweiler Foundation. The Centre for Trials Research, Cardiff University receives infrastructure funding from Health and Care Research Wales. This work was also supported by Medical Research Council (UK) Grant MR/L010305/1. Funding was also provided by grant R01DA045049 (PI Griffin) through the National Institute of Drug Abuse.

References

Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., Kong, M.: Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: a comparative study. Biometric. J. **59**(5), 967–985 (2017)

Agresti, A.: An Introduction to Categorical Data Analysis. Wiley, Hoboken (2018)

Ali, M.S., Groenwold, R.H., Klungel, O.H.: Best (but oft-forgotten) practices: propensity score methods in clinical nutrition research. Am. J. Clin. Nutr. **104**(2), 247–258 (2016)

Altman, D.G., Bland, J.M.: Treatment allocation in controlled trials: Why randomise? BMJ **318**(7192), 1209–1209 (1999)

Austin, P.C.: A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat. Med. **27**(12), 2037–2049 (2008)

Austin, P.C.: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat. Med. **28**(25), 3083–3107 (2009)

Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar. Behav. Res. **46**(3), 399–424 (2011)

Austin, P.C., Mamdani, M.M.: A comparison of propensity score methods: a case-study estimating the effectiveness of post-ami statin use. Stat. Med. **25**(12), 2084–2106 (2006)

Austin, P.C., Stuart, E.A.: Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Stat. Med. **34**(28), 3661–3679 (2015)

Austin, P.C., Grootendorst, P., Anderson, G.M.: A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. Stat. Med. **26**(4), 734–753 (2007)

Bang, H., Robins, J.M.: Doubly robust estimation in missing data and causal inference models. Biometrics **61**(4), 962–973 (2005)

Bergstra, S.A., Sepriano, A., Ramiro, S., Landewé, R.: Three handy tips and a practical guide to improve your propensity score models. RMD Open **5**(1), e000953 (2019)

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., Stürmer, T.: Variable selection for propensity score models. Am. J. Epidemiol. **163**(12), 1149–1156 (2006)

Burgette, L.F., McCaffrey, D.F., Griffin, B.A.: Propensity score estimation with boosted regression. In: Propensity Score Analysis: Fundamentals, Developments and Extensions. Guilford Publications Inc, New York (2015)
Caliendo, M., Kopeinig, S.: Some practical guidance for the implementation of propensity score matching. J. Econ. Surv. **22**(1), 31–72 (2008)
Chattopadhyay, A., Hase, C.H., Zubizarreta, J.R.: Balancing versus modeling approaches to weighting in practice. Stat. Med. **39**(24), 3227–3254 (2020)
Choi, B.Y., Wang, C.P., Michalek, J., Gelfond, J.: Power comparison for propensity score methods. Comput. Statist. **34**(2), 743–761 (2019)
Cox, D.R., Cox, D.R.: Planning of Experiments, vol. 20. Wiley, New York (1958)
D'Agostino, R.B., Jr.: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat. Med. **17**(19), 2265–2281 (1998)
Dennis, M.L., Titus, J.C., White, M.K., Unsicker, J.I., Hodgkins, D.: Global Appraisal of Individual Needs: Administration Guide for The Gain and Related Measures. Chestnut Health Systems, Bloomington (2003)
Elze, M.C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G.W., Pocock, S.J.: Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. J. Am. Coll. Cardiol. **69**(3), 345–357 (2017)
Franklin, J.M., Rassen, J.A., Ackermann, D., Bartels, D.B., Schneeweiss, S.: Metrics for covariate balance in cohort studies of causal effects. Stat. Med. **33**(10), 1685–1699 (2014)
Gail, M.H., Green, S.B.: Critical values for the one-sided two-sample kolmogorov-smirnov statistic. J. Am. Stat. Assoc. **71**(355), 757–760 (1976)
Garrido, M.M., Kelley, A.S., Paris, J., Roza, K., Meier, D.E., Morrison, R.S., Aldridge, M.D.: Methods for constructing and assessing propensity scores. Health Serv. Res. **49**(5), 1701–1720 (2014)
Godley, S.H.: The adolescent community reinforcement approach for adolescent cannabis users. Vol. 4. US Department of Health and Human Services (2001). https://books.google.co.uk/books?hl=en&lr=&id=lCdOAQAAIAAJ&oi=fnd&pg=PP13&dq=Godley+S.H.:+The+adolescent+community+reinforcement+approach+for+adolescent+cannabis+users+US+Department+of+Health+and+Human+Services+etal.(2017)
Grant, S., Hunter, S.B., Pedersen, E.R., Griffin, B.A.: Practical factors determining adolescent substance use treatment settings: results from four online stakeholder panels. J. Subst. Abuse Treat. **109**, 34–40 (2020)
Griffin, B.A., McCaffrey, D.F., Ramchand, R., Hunter, S.B., Booth, M.S.: Assessing the sensitivity of treatment effect estimates to differential follow-up rates: implications for translational research. Health Serv. Outcomes Res. Method **12**(2), 84–103 (2012)
Griffin, B.A., Ramchand, R., Almirall, D., Slaughter, M.E., Burgette, L.F., McCaffery, D.F.: Estimating the causal effects of cumulative treatment episodes for adolescents using marginal structural models and inverse probability of treatment weighting. Drug Alcohol Depend. **136**, 69–78 (2014)
Griffin, B.A., McCaffrey, D.F., Almirall, D., Burgette, L.F., Setodji, C.M.: Chasing balance and other recommendations for improving nonparametric propensity score models. J. Causal Inference **5**(2), (2017). https://www.degruyter.com/document/doi/10.1515/jci-2015-0026/html
Griffin, B.A., Ayer, L., Pane, J., Vegetabile, B., Burgette, L., McCaffrey, D., Coffman, D.L., Cefalu, M., Funk, R., Godley, M.D.: Expanding outcomes when considering the relative effectiveness of two evidence-based outpatient treatment programs for adolescents. J. Subst. Abuse Treat. **118**, 108075 (2020)
Hainmueller, J.: Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. Polit. Anal. **20**(1), 25–46 (2012)
Harder, V.S., Stuart, E.A., Anthony, J.C.: Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol. Methods **15**(3), 234 (2010)
Hernán, M.Á, Brumback, B., Robins, J.M.: Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 561–570 (2000). https://www.jstor.org/stable/3703998?casa_token=UoTMvudmO1YAAAAA%3AqxfNrvRxRo8KXoY4mulfuzqTtt9edsXioEqm9Cr35yPz_YcQTik6m9cgXxmlCVSIewXN3Tf8Ku6Xrb6Br5qXYumYwhd9_iGtIoUDD5uouYzopcLtw&seq=1
Hirano, K., Imbens, G.W.: Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv. Outcomes Res. Method. **2**(3–4), 259–278 (2001)
Ho, D.E., Imai, K., King, G., Stuart, E.A.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit. Anal. **15**(3), 199–236 (2007)
Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**(396), 945–960 (1986)

Huang, M., Vegetabile, B., Burgette, L., Setodji, C., Griffin, B.A.: Balancing higher moments matters for causal estimation: further context for the results of Setodji et al. (2017). (2021). arXiv:2107.03922

Imai, K., Ratkovic, M.: Covariate balancing propensity score. J. R. Statist. Soc. Ser. B (Statist. Methodol.) **76**(1), 243–263 (2014)

Kang, J.D., Schafer, J.L., et al.: Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat. Sci. **22**(4), 523–539 (2007)

King, G., Nielsen, R.: Why propensity scores should not be used for matching. Polit. Anal. **27**(4), 435–454 (2019)

Lee, J., Little, T.D.: A practical guide to propensity score analysis for applied clinical research. Behav. Res. Ther. **98**, 76–90 (2017)

Leite, Walter: Practical propensity score methods using R. Sage Publications, Thousand Oaks (2016)

Li, F., Morgan, K.L., Zaslavsky, A.M.: Balancing covariates via propensity score weighting. J. Am. Stat. Assoc. **113**(521), 390–400 (2018)

Mao, H., Li, L., Greene, T.: Propensity score weighting analysis and treatment effect discovery. Stat. Methods Med. Res. **28**(8), 2439–2454 (2019)

Markoulidakis, A., Holmans, P., Pallmann, P., Busse, M., Griffin, B.A.: How balance and sample size impact bias in the estimation of causal treatment effects: a simulation study. arXiv:2107.09009 (2021)

Markoulidakis, A., Holmans, P., Pallmann, P., Busse, M., Griffin, B.A.: CoBWeb: a user-friendly web application to estimate causal treatment effects from observational data using multiple algorithms. arXiv:2112.05035 (2021)

McCaffrey, D.F., Ridgeway, G., Morral, A.R.: Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol. Methods **9**(4), 403 (2004)

McCaffrey, D.F., Griffin, B.A., Almirall, D., Slaughter, M.E., Ramchand, R., Burgette, L.F.: A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat. Med. **32**(19), 3388–3414 (2013)

Mlcoch, T., Hrnciarova, T., Tuzil, J., Zadak, J., Marian, M., Dolezal, T.: Propensity score weighting using overlap weights: a new method applied to regorafenib clinical data and a cost-effectiveness analysis. Value Health **22**(12), 1370–1377 (2019)

Myers, J.A., Louis, T.A.: Regression adjustment and stratification by Propensty score in treatment effect estimation (2010). https://biostats.bepress.com/jhubiostat/paper203/

Olmos, A., Govindasamy, P.: A practical guide for using propensity score weighting in R. Pract. Assess. Res. Eval. **20**(1), 13 (2015)

Pane, J.D., Griffin, B.A., Burgette, L.F., McCaffrey, D.F.: Ovtool-omitted variable tool, v.1.0.3 (2021). https://cran.r-project.org/web/packages/OVtool/index.html

Pirracchio, R., Petersen, M.L., van der Laan, M.: Improving propensity score estimators' robustness to model misspecification using super learner. Am. J. Epidemiol. **181**(2), 108–119 (2015)

Posner, M.A., Ash, A.S.: Comparing weighting methods in propensity score analysis. Unpublished working paper, Columbia University (2012)

Ramchand, R., Griffin, B.A., Suttorp, M., Harris, K.M., Morral, A.: Using a cross-study design to assess the efficacy of motivational enhancement therapy-cognitive behavioral therapy 5 (met/cbt5) in treating adolescents with cannabis-related disorders. J. Stud. Alcohol Drugs **72**(3), 380–389 (2011)

Ramchand, R., Griffin, B.A., Slaughter, M.E., Almirall, D., McCaffrey, D.F.: Do improvements in substance use and mental health symptoms during treatment translate to long-term outcomes in the opposite domain? J. Subst. Abuse Treat. **47**(5), 339–346 (2014)

Ramchand, R., Griffin, B.A., Hunter, S.B., Booth, M.S., McCaffrey, D.F.: Provision of mental health services as a quality indicator for adolescent substance abuse treatment facilities. Psychiatr. Serv. **66**(1), 41–48 (2015)

Ratkovic, M., Imai, K., Fong, C.: Package 'CBPS'. Maintainer Marc Ratkovic (2013). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.409.3744&rep=rep1&type=pdf

Ridgeway, G.: The state of boosting. Comput. Sci. Statist. 172–181 (1999). http://www.planchet.net/EXT/ISFA/1226.nsf/0/a29acbd26d902d6fc125822a0031c09b/$FILE/boosting.pdf

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., Griffin, B.A.: Toolkit for Weighting and Analysis of Nonequivalent Groups: A Tutorial for the Twang Package. RAND Corporation, Santa Monica, CA (2017)

Robins, J.M., Hernan, M.A., Brumback, B.: Marginal structural models and causal inference in epidemiology. Epidemiology **11**(5), 550–560 (2000). https://journals.lww.com/epidem/Fulltext/2000/09000/Marginal_Structural_Models_and_Causal_Inference_in.11.aspx/?casa_token=PXWQlSYRV5IAAAAA:5plsLMdPpYgRMn_rgcpERV9TZl_MQLi8SxPHk26Z6YsCEu86bOm5TBdiiRhs8UuHiKwIrrxfkWxJNnerWN8-fZ_xn0or

Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**(5), 688 (1974)
Sampl, S., Kadden, R.: Motivational enhancement therapy and cognitive behavioral therapy for adolescent cannabis users: 5 sessions. In: Cannabis Youth Treatment (CYT) Series, vol. 1. (2001). https://eric.ed.gov/?id=ED478681
Schuler, M.S., Griffin, B.A., Ramchand, R., Almirall, D., McCaffrey, D.F.: Effectiveness of treatment for adolescent substance use: is biological drug testing sufficient? J. Stud. Alcohol Drugs **75**(2), 358–370 (2014)
Setodji, C.M., McCaffrey, D.F., Burgette, L.F., Almirall, D., Griffin, B.A.: The right tool for the job: choosing between covariate balancing and generalized boosted model propensity scores. Epidemiology **28**(6), 802 (2017)
Setoguchi, S., Schneeweiss, S., Brookhart, M.A., Glynn, R.J., Cook, E.F.: Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol. Drug Saf. **17**(6), 546–555 (2008)
Shook-Sa, B.E., Hudgens, M.G.: Power and sample size for observational studies of point exposure effects. Biometrics **78**(1), 388–398 (2022). https://onlinelibrary.wiley.com/doi/full/10.1111/biom.13405?casa_token=f4dzKzz5puwAAAAA%3AZDH8hLzSJIh1jAo5g6bsk8Ar4RKBcppKfVVGSDTvFljG01qveaaMQyeyyfmfVKWDfJxsd9SPJG4bTA
Stuart, E.A.: Matching methods for causal inference: a review and a look forward. Stat. Sci. **25**(1), 1–21 (2010). https://doi.org/10.1214/09-STS313
Stuart, E.A., Lee, B.K., Leacy, F.P.: Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J. Clin. Epidemiol. **66**(8), S84–S90 (2013)
VanderWeele, T.J., Ding, P.: Sensitivity analysis in observational research: introducing the E-value. Ann. Intern. Med. **167**(4), 268–274 (2017)
Vegetabile, B.: entbal: An Alternative Implementation of Entropy Balancing Weights For Estimating Causal Effects. GitHub, GitHub Repository (2021). https://github.com/bvegetabile/entbal
Wright, R.E.: Logistic Regression. American Psychological Association, Washington, D.C. (1995)
Wyss, R., Ellis, A.R., Brookhart, M.A., Girman, C.J., Jonsson Funk, M., LoCasale, R., Stürmer, T.: The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bcart, and the covariate-balancing propensity score. Am. J. Epidemiol. **180**(6), 645–655 (2014)
Xie, Y., Zhu, Y., Cotton, C.A., Wu, P.: A model averaging approach for estimating propensity scores by optimizing balance. Stat. Methods Med. Res. **28**(1), 84–101 (2019)
Zhao, Q., et al.: Covariate balancing propensity score by tailored loss functions. Ann. Stat. **47**(2), 965–993 (2019)
Zhang, Z., Kim, H.J., Lonjon, G., Zhu, Y., et al.: Balance diagnostics after propensity score matching. Ann. Trans. Med. **7**(1) (2019). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351359/
Zhao, Q., Percival, D.: Entropy balancing is doubly robust. J. Causal Inference **5**(1) (2017). https://www.degruyter.com/document/doi/10.1515/jci-2016-0010/html

Authors and Affiliations

Andreas Markoulidakis[1,2] · Khadijeh Taiyari[2] · Peter Holmans[3] · Philip Pallmann[2] · Monica Busse[2] · Mark D. Godley[4] · Beth Ann Griffin[5]

1 Centre for Trials Research, Cardiff University, Cardiff, Wales, UK

2 School of Medicine, Cardiff University, Cardiff, Wales, UK

3 Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, Wales, UK

4 Chestnut Health Systems, Normal, IL, USA

5 RAND Corporation, Arlington, VA, USA