A tutorial comparing different covariate balancing methods with an application evaluating the causal effect of exercise on the progression of Huntington's Disease

Andreas Markoulidakis[a,*], Khadijeh Taiyari[c], Peter Holmans[b], Philip Pallmann[c]

Monica Busse − Morris[c], Beth Ann Griffin[d]

[a]School of Medicine, Cardiff University, Cardiff, UK
[b]Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK
[c]Centre for Trials Research, Cardiff University, Cardiff, UK
[d]RAND Corporation, Arlington, VA, USA

[*]$MarkoulidakisA@cardiff.ac.uk$

Randomized controlled trials are the gold standard for measuring the causal effects of treatments on clinical outcomes. However, randomized trials are not always feasible, and causal treatment effects must, therefore, often be inferred from observational data. Observational study designs do not allow conclusions about causal relationships to be drawn unless statistical techniques are used to account for the imbalance of confounders across groups while key assumptions hold. Propensity score (PS) and balance weighting are two useful techniques that aim to reduce the imbalances between treatment groups by weighting the groups to look alike on the observed confounders. There are many methods available to estimate PS and balancing weights. However, it is unclear a priori which will achieve the best trade-off between covariate balance and effective sample size. Weighted analyses are further complicated by small studies with limited sample sizes, which is common when studying rare diseases. To address these issues, we present a step-by-step guide to covariate balancing strategies, including how to evaluate overlap, obtain estimates of PS and balancing weights, check for covariate balance, and assess sensitivity to unobserved confounding. We compare the performance of a number of commonly used estimation methods on a synthetic data set based on the Physical Activity and Exercise Outcomes in Huntington Disease (PACE-HD) study, which explored whether enhanced physical activity affects the progression and severity of the disease. We provide general guidelines for the choice of method for estimation of PS and balancing weights, interpretation, and sensitivity analysis of results. We also present R code for implementing the different methods and assessing balance.

1. Introduction

In a randomized controlled trial (RCT), the assignment to treatment[1] 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 & 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

---

[1] An intervention in lifestyle is a special form of treatment.

conditions of the individual. These differences between the groups will introduce bias in to the estimated effects of a treatment. For example, the choice to exercise among individuals with a rare 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 pretreatment levels of disease severity. Our motivating example is based on the Physical Activity and Exercise Outcomes in Huntington Disease (PACE-HD) study (Drew et al., 2019), a trial-within-observational cohort study where individuals in treatment group of the RCT receive an intervention of supervised physical activity, and the goal is to explore the impact of exercise on the progression and severity of HD versus controls from the RCT as well as a cohort of controls for a larger observational study.

Estimation of accurate causal treatment effects (Holland, 1986) is the main goal of many observational studies. 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 & Rubin, 1983). As a consequence, we design longitudinal, observational studies, where two (or more) groups receive different treatments overtime, 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.

The propensity score (PS) (Rosenbaum & Rubin, 1983) is the probability of an individual's allocation to the treatment group, given their observed baseline (pretreatment) characteristics. The PS can be used to create comparable treatment groups 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. Here we consider only PS weighting (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 & Govindasamy, 2015; Posner & Ash, 2012). In addition to PS weights, we consider the closely related balancing weights via entropy balancing (Hainmueller, 2012). Entropy Balance computes the balancing weight directly, as opposed to traditional PS algorithms, which at first compute the PS and then transform it to weights. Thus, we are focusing only on PS and balancing weights methods to allow for a direct comparison of the performance of different algorithms.

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 & Ratkovic, 2014). We will also consider the use of balancing weights via Entropy Balance (EB) (Hainmueller, 2012) and provide a summary of the advantages and disadvantages of each approach. The performance of the

methods will be illustrated by applying them to a synthetic dataset with a structure that is characteristic of many studies in rare disease populations with small sample sizes.

The rest of the article is organised as follows. *Section* 2 summarizes the six key steps needed to estimate causal treatment effects of a treatment in observational studies. Then, the PS and EB estimation algorithms we use are described in *section* 3, and the measures of performance of the different balancing methods are briefly discussed in *section* 4. *Section* 5 describes the data example and *section* 6 uses the data to walk through the 6 key steps needed to estimate causal effects. *Section* 7 concludes. The code for each step of *section* 6 is provided in 8.

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 & 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)

↓

Step 2

Assess sample for any obvious overlap concerns and adjust sample as needed

↓

Step 3

Estimate the PS or balancing weight, ideally using multiple methods

↓

Step 4

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

↓

Step 5

Model the outcome and estimate the causal treatment effect

↓

Step 6

Assess sensitivity of the results to unobserved confounding

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

In order to define the commonly used causal treatment effects, we first introduce potential outcomes notation first introduced by Rubin (Griffin et al., 2014). For each individual, we define Y_{1i} to denote their potential outcome under treatment and Y_{0i} to denote their potential outcome under control for $i = 1, \ldots N$. While these potential outcomes exist in theory for all individuals in our study, we only get to observe one potential outcome for each individual in the study. Namely, we observed the outcome for which treatment they were observed to receive. We define our treatment indicator as T_i where values of 1 denote that individual 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 are:

– the *Average Treatment Effect on the Entire population*

$$ATE = E[Y_1] - E[Y_0], \quad (1)$$

– the *Average Treatment Effect on the Treated population*

$$ATT = E[Y_1|T = 1] - E[Y_0|T = 1], \quad (2)$$

– the *Average Treatment Effect on the Control population*

$$ATC = E[Y_1|T = 0] - E[Y_0|T = 0]. \quad (3)$$

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 the PACE-HD study, the ATE would measure the effect of exercise for all people with HD like those enrolled in the study, while ATT and ATC would quantify the effect of exercise for individuals with HD who are and are not exercising, respectively. Since individuals who exercise will tend to be different from those who do not (e.g., healthier), the populations to which ATT and ATC generalize will be different.

2. **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 & Rubin, 1983) — *stable unit treatment value assumption* and *strong ignorability*. SUTVA implies that the distribution of potential outcomes for each individual is independent of the potential outcomes of other individuals (Cox & Cox, 1958). Strong ignorability includes two key parts related to the PS (or treatment assignment mechanism), defined here as $P(T_i = 1|X_i)$ where 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 ($0 < P(T_i = 1|X_i) < 1$) (Rosenbaum & 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, sd, max, min) 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 shown in *Figure* 1.

If there are ranges of covariate values that only occur in one group, one might have to exclude some individuals, such that the two groups are adequately overlapped, or consider other estimands, like the *Average Treatment Effect on the 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.

We use *figure* 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 when a study lacks overlap. It is apparent from *figure* 1 that the support of the control group (blue 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 or by estimating the treatment effect only for the region of overlap (ATO) (Li et al., 2018).
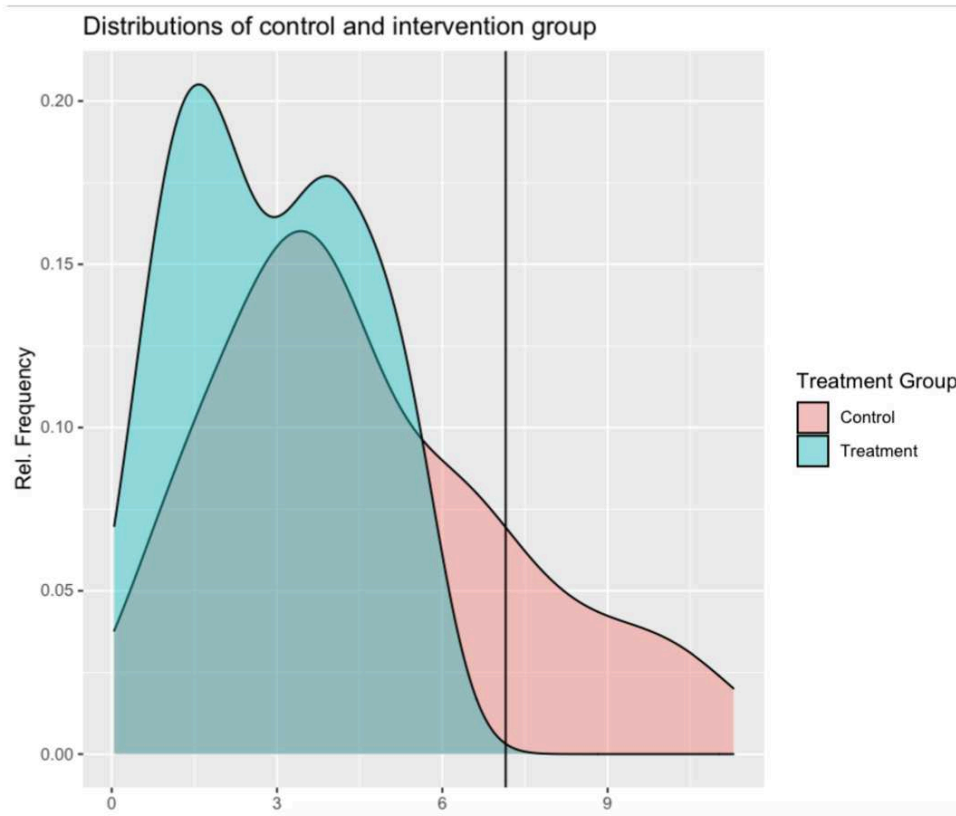
Figure 1: Density plot of control and treatment group, where there is lack of overlap in the support of the two groups. The data used for this plot are artificial to demonstrate the issue.

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.

3. **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, (Griffin et al., 2017; Setodji et al., 2017; Setoguchi et al., 2008)). Under different settings, different methods perform better, and this depends on the structure of the given dataset — 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-parametric). 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).

4. **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 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 & Green, 1976; Setodji et al., 2017; Setoguchi et al., 2008). Both metrics are explained in detail in Section 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 weight with the lowest reduction to the ESS. The ESS is also explained in more detail in Section 4.

5. **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, researchers 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], a study will not be able to robustly estimate the causal effect of the treatment of interest. It is important that researchers understand when this is not happening, the study can only be used to examine associations since the findings will be less robust and must be caveated as such.

Assuming one does have adequate balance and sufficient 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:

$$AT\_ = \frac{\sum_{j \in C_1} w_i Y_i^{obs}}{\sum_{i \in C_1} w_i} - \frac{\sum_{j \in C_0} w_i Y_i^{obs}}{\sum_{i \in C_0} w_i}$$

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). When using PS weights, an estimated treatment effect that comes from a regression model that includes all observed confounders along with the PS weights is a doubly robust treatment effect (Bang & Robins, 2005; Kang et al., 2007). 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). 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.

6. **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

---

[2] Two, or more, groups are considered to be perfectly balance as soon as SMD and KS-Statistiv values are below 0.1 threshold.

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 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).

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 $K-$dimensional column vector of observed pre-treatment covariates X_i. The PS is defined as the conditional probability of receiving the treatment given the covariates X_i, i.e. $\pi(X_i) = Pr(T_i = 1|X_i)$. From (Rosenbaum & Rubin, 1983), the ignorability of treatment assignment says that the treatment assignment is ignorable given the (true) propensity score $\pi(X_i)$. This implies that the unbiased estimation of treatment effect is possible by conditioning on the PS alone instead of the entire covariate vector X_i. However, in observational studies the PS must be estimated from the dataset. Normally, one assumes a parametric PS model $\pi_\beta(X_i)$. That is $Pr(T_i = 1|X_i) = \pi_\beta(X_i)$, where $\beta \in \delta$ is an $L-$ dimensional column vector of unknown parameters.

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

3.1 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:

$$\pi_\beta(X_i) = \frac{exp(X_i^T \beta)}{1 + exp(X_i^T \beta)}$$

An implementation of LR for PS estimation in R software is discussed in (Olmos & Govindasamy, 2015)o. 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. It is nearly impossible to know the best way to specify the right hand side of the LR model.

3.2 Covariate Balance 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 & Ratkovic, 2014).

The authors used the covariate balancing property of parametric models by employing inverse PS weighting:

$$E\left(\frac{T_i f(X_i)}{\pi_\beta(X_i)} - \frac{(1 - T_i) f(X_i)}{1 - \pi_\beta(X_i)}\right) = 0$$

where $f(X_i)$ is an M−dimensional vector-values 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.

This method has the advantage of being robust to mild model misspecification 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 datasets and improve the accuracy of estimated treatment effects over parametric models even if there is no misspecification (Choi et al., 2019; Imai & 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.

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)), 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.

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

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. a GBM does not capture 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 TWANG package in R implements optimization of the PS model using GBM optimized on best balance rather than best model fit (Ridgeway et al., 2020).

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).

3.4 Entropy Balance

Entropy balance is a method that aims to estimate the weights directly rather than the PS of the individuals. The method promises to achieve exact balance on as many moments as defined by the user.

Assuming that one is interested to estimate the ATT, the quantity that is tricky to estimate is $E[Y_{i|T0}|T_i = 1]$, since these values are not observed. Thus, an estimation of the above quantity is

$$E[\widehat{Y_0|T} = 1] = \frac{\sum_{C_0} w_i^0 Y_i^{obs}}{\sum_{i=1}^{n_0} w_i^0},$$

where w_i are the balance weights and need to be estimated. Entropy Balance method calculates weights through a reweighting scheme (until adequate balance in the pre-advised moments is achieved), while at the same time attempting to match the first k moments of thedistributions of the two groups (k is defined by the user). Entropy balance re-weighting schemes can be considered as generalizations of the traditionally used IPW (Hainmueller, 2012). However, the weights here are not defined by the PS, thus a direct estimation of PS is not feasible.

Entropy balance can be implemented in R, with the package *entbal* (*Vegetabile BG. A Tutorial on Entropy Balancing and Weighted Estimation of Causal Effects: A Guide to the 'entbal' Package in R. R package. RAND Tool. RAND Corporation. Under review.*), however, the package is designed to match only the first moment.

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

$$w_i = T_i \frac{1}{\pi(x_i)} + (1 - T_i) \frac{1}{1 - \pi(x_i)}, \qquad for\ ATE \qquad (4)$$

$$w_i = T_i + (1 - T_i) \frac{\pi(x_i)}{1 - \pi(x_i)}, \qquad for\ ATT \qquad (5)$$

$$w_i = T_i \frac{1 - \pi(x_i)}{\pi(x_i)} + (1 - T_i), \qquad for \ ATC \qquad (6)$$

Logistic regression, *GBM* and *CBPS* produces PS, which we transform to balancing weights through the formula described above, whereas *Entropy Balance* estimates the weights directly. For LR we will use the $glm()$ function, which is part of the $stats$ package in R. For *GBM* we will use the $twang$ package (Ridgeway et al., 2020), while for *CBPS*, the $CBPS$ package (Fong et al., 2019) (which is accompanied with $COBALT$ package (Greifer, 2020) for plots and diagnostics) will be used. Finally, for *Entropy Balance* we will use the package $entbal$.

4. Measures to Evaluate Balance

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

4.1 Standardized Mean Difference

SMD (Austin, 2009; Franklin et al., 2014) 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:

$$SMD = \frac{\bar{X}_{treatment} - \bar{X}_{control}}{\sqrt{\dfrac{(n_{treatment} - 1)sd^2_{treatment} + (n_{control} - 1)sd^2_{control}}{n_{treatment} + n_{control} - 2}}},$$

where $X_{treatment}$ and $X_{control}$ are the sample means, $sd_{treatment}$ and $sd_{control}$ are the standard deviations and $n_{treatment}$ and $n_{control}$ are the sample sizes of treatment and control group, respectively. For ATT, the standard deviation used would be $sd_{treatment}$, while for ATC, in the denominator one should consider $sd_{control}$.

SMD values lie between 0 and 1, and the lower the value, the better the balance achieved. In the literature, 0.1 is recommended as a threshold to define groups as balanced. By definition, it quantifies the similarity of the means of the two groups for each variable, thus it expresses how close the two mean values are.

4.2 Kolmogorov-Smirnov Statistic

KS is a statistical test (Gail & Green, 1976) which checks the hypothesis that the two samples are from the same distribution. Its test statistic is formally defined as

$$KS = \max_z \left| F^{treatment}_{emp}(z) - F^{control}_{emp}(z) \right|,$$

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 $x_1, x_2, \ldots, x_n$, is:

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

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. The is no clear guidance on what is the best threshold for the KS but values over 0.1 would be consider notable large and thus, we propose to use 0.1 as the threshold for balance for the KS as well as the SMD.

4.3 Effective Sample Size

Given the weights of each group, the ESS (Ridgeway et al., 2017) is defined as:

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

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 needs 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.

5. Study Data

The data example used in this tutorial is based on the PACE-HD study (Drew et al., 2019). It is an observational cohort study with a nested randomized trial of long term (e.g. 12 months) exercise compared to usual physical activity on progression and severity of Huntington's Disease (HD). Data consist of 120 people with a confirmed genetic diagnosis of HD that satisfy inclusion-exclusion criteria were included in the PACE-HD study. The allocation mechanism between the wider control and treatment group is not random. Therefore, for the purposes of this tutorial, we will utilize the part of the control group that does not contain the control group of the nested randomized trial.

There are baseline data for 111 individuals, 58 of which are allocated to the control group, and the rest being split between control (27) and treatment (26) groups through a randomization process. Due to the small size of the available dataset, care is required to select both the method and covariates to use for balancing. In this tutorial, we will control for the 3 main confounders $VO2max$, $cUHDRS\_Y1$ and $6'\ Walk\ Distance$ (for formal definitions of these covariates see [9]). In brief, $cUHDRS\_Y1$ denotes the baseline severity of HD for each individual while $VO2max$ and $6'\ Walk\ Distance$ measure the physical fitness levels of each individuals. These confounders were chosen based on scientific advice from the study team — dealing with variable selection for PS and outcome model is beyond the goal of this current tutorial (Brookhart et al., 2006; Hirano & Imbens, 2001). Our key outcome is $cUHDRS\_Y12$ measured 12-months after intake.

Since it is not feasible to publish the original data (for identification reasons), we simulated synthetic data (using the same mean and covariance matrix with the original data) to demonstrate our step-by-step procedure to obtain an estimation of causal treatment effect of exercise in *section 6*. It is common in observational studies to have a mixture of binary and continuous covariates, thus it is often questionable how to simulate binary and continuous covariates at the same time to preserve the correlation across the covariates of different types. We will utilize the R package *synthpop* (Nowok et al., 2019), which is based on (Nowok et al., 2016) and generates synthetic data based on the structure of an original dataset.
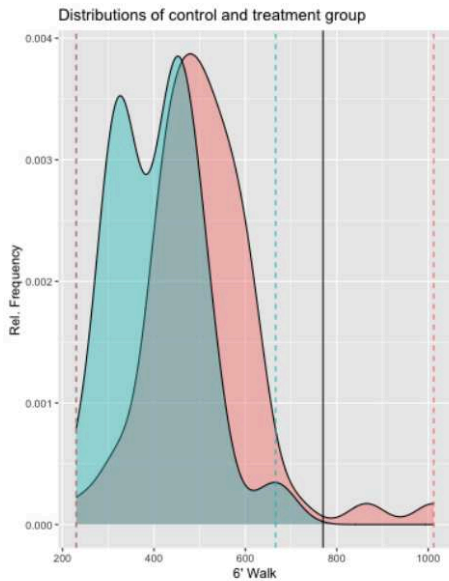
6. Obtaining PS Weighting estimates and assessing their performance

## 6.1 Choose which estimand one is interested in (ATE, ATT, ATC)

Since we would expect individuals who are exercising to be healthier on average than those who do not, we felt that it would be most meaningful for the PACE-HD study to estimate the ATT of exercise among HD individuals like those who were selecting to exercise. Thus, we estimate ATT in our illustration.

6.2 Assess Sample for any obvious Overlap Concerns and Adjust Sample size as Needed

After step 1 we continue with assessment of the overlap between the control and treatment groups. To do so, we create a table of minimum and maximum value per group, per variable (*table* 1), as well as distribution plot (*figure* 2), which provides a visualisation of the overlap, such that any regions of concern can be detected. The code to implement this analysis may be found in 8.1.
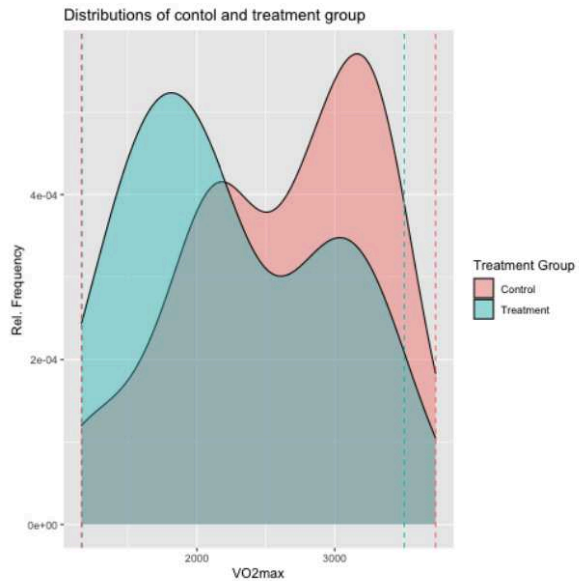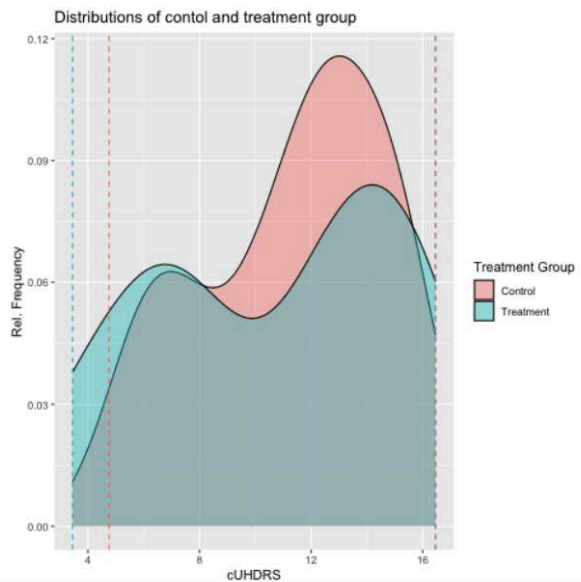
Figure 2: Density plots of $6'$ Walk, V O2max and cUHDRS (with light blue is the control group and with light red the treatment group). The vertical line in $6'$ Walk plot indicates the point beyond of which, the two densities do not overlap.

Table 1: Table of mean, standard deviation, minimum and maximum value per variable, per group. Based on min/max values per group, it seems that there is a lack of overlapping on 6′ Walk for the two group, which will be checked on the distribution plots.

	Control Group				Treatment Group			
	Mean	**SD**	**min**	**max**	**Mean**	**SD**	**min**	**max**
6' Walk	509.48	122.41	230.00	1012.00	407.01	95.74	230.00	666.20
cUHDRS_Y1	11.28	3.30	4.76	16.46	10.65	4.36	3.45	16.46
VO2max	2615.62	685.61	1163.19	3730.87	2243.57	710.57	1166.46	3504.06

Table 1: Table of mean, standard deviation, minimum and maximum value per variable, per group. Based on min/max values per group, it seems that there is a lack of overlapping on $6'$ Walk for the two group, which will be checked on the distribution plots.

From *table* 1, it seems that there is a lack of overlap on the maximum value of $6'\ Walk$ for the two groups, since for control group the maximum value is 1012 metres, while for treatment it is 666.2 metres. This is clear in *figure* 2, where the $6'\ Walk$ plot shows a lack of overlap in the right tail of the distribution of the control group. For this reason, we remove these two outlying individuals from the analysis to ensure adequate overlap in our two groups. After the removal of the two outliers, the distribution plots are as depicted in *figure 3*. This corrects the overlapping issue, although we report balance measures for both datasets, with and without outliers, to check whether outliers reduce the resulting balance.

(a) 6′ Walk
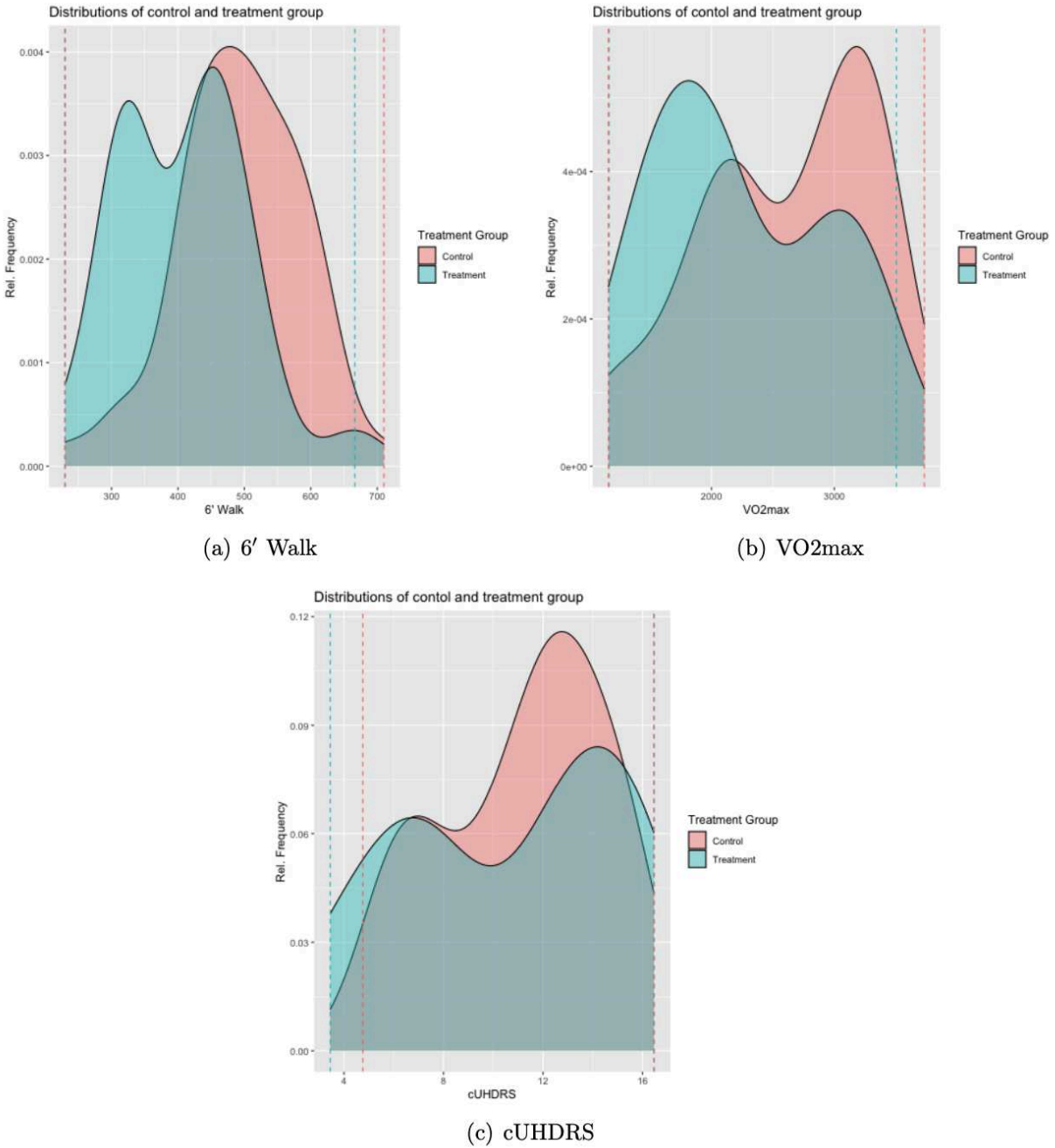
(b) VO2max

(c) cUHDRS

Figure 3: Density plots of 6′ Walk, V O2max and cUHDRS, without the outliers individuals.

6.3 Estimation of Propensity Scores or Balancing Weights, ideally using Multiple Methods

Next, we estimate balancing weights using the four methods presented in Section 3. Here, we are interested in the estimation of ATT, thus the relevant formulae for estimation of balancing weights from PS for LR, GBM and CBPS — for estimation of ATE and/or ATC, one should use formulas ([ATE_weights]) and ([ATC_weights]) to transform PS to balancing weights, respectively, and

shift the relevant option in *EntBal* function. The code to implement this analysis can be found in 8.2.

6.4 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. The code to compute *SMD, KS and ESS* for each balancing method is described in 8.3. *Table* 2 reports the SMD and KS values per method, per variable (rounded to 2 decimal places), as well as the ESS per method.

Table 2: Standardized Mean Difference (SMD), Kolmogorov-Smirnov Statistic (KS) and Effective Sample Size (ESS) per estimation method. For SMD and KS 0.1 is considered as a widely accepted threshold, while ESS should be as close to unweighted as possible.

SMD								
	Unweighted	LR	GBM	GBM	CBPS	EB #1	EB #2	EB #3
SIX MIN Walk	0.91	0.07	0.06	0.04	0.04	0.00	0.00	0.00
VO2max	0.52	0.02	0.13	0.17	0.03	0.00	0.00	0.00
cUHDRS_Y1	0.12	0.11	0.05	0.04	0.01	0.00	0.00	0.00
mean	0.52	0.07	0.08	0.08	0.03	0.00	0.00	0.00
max	0.91	0.11	0.13	0.17	0.04	0.00	0.00	0.00
KS Statistic								
	Unweighted	LR	GBM	GBM	CBPS	EB #1	EB #2	EB #3
SIX MIN Walk	0.43	0.15	0.18	0.17	0.19	0.18	0.21	0.27
VO2max	0.37	0.10	0.14	0.16	0.11	0.10	0.12	0.17
cUHDRS_Y1	0.21	0.17	0.12	0.12	0.14	0.13	0.12	0.15
mean	0.34	0.14	0.15	0.15	0.15	0.14	0.15	0.20
max	0.43	0.17	0.18	0.17	0.19	0.18	0.21	0.27
ESS								
	Unweighted	LR	GBM	GBM	CBPS	EB #1	EB #2	EB #3
	56.0	20.2	15.0	16.5	23.2	21.4	18.9	13.1

Table 2: Standardized Mean Difference (SMD), Kolmogorov-Smirnov Statistic (KS) and Effective Sample Size (ESS) per estimation method. For SMD and KS 0.1 is considered as a widely accepted threshold, while ESS should be as close to unweighted as possible.

For SMD and KS statistic, 0.1 is a widely accepted threshold — any difference under this threshold is not considered as evidence for imbalance. As observed in *table* 2, *EB* achieves the best balance in terms of SMD (absolutely 0 for all covariates), followed closely by *CBPS*, which reports mean SMD 0.03 with a maximum value of 0.04. Both methods perform well in terms of

SMD, especially when considering the unweighted SMD of $6' \, Walk$ of 0.91 — which is considered as an extreme case of imbalance.

In terms of KS statistic, no method achieves an acceptable level of balance. LR and $EB\#1$ give the lowest values, although the other methods are similar. This lack of balance may be explained by the high initial KS and the small sample sizes per group (58 in the control group and 26 in the treatment group). This is also reflected in the big losses of effective sample size for all methods, with $CBPS$ giving the highest ESS (losing over 50% compared to the unweighted analysis). Losses of 50% are relatively common for analyses with meaningful differences between the groups being compared and such reductions in sample size should not be shocking. The impact is more difficult for small studies.

EB (controlling for the 1st moment, and CBPS seem to perform identically, however, given that the results are rounded to 2 decimal places, we pick EB as the method with the best balance (the SMD scores are exactly 0). Thus, we apply the outcome and sensitivity analysis to EB #1.

6.5 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 obtained in previous steps. To do so, we fit a weighted linear regression on $cUHDRS_{year2}$ (this is the $cUHDRS$ at the end of the treatment period) considering as predictors the $6' \, Walk$, $cUHDRS\_Y1$ and $VO2max$ at the baseline, and the treatment status. The coefficient of treatment status represents the estimand of interest.

Table 3: Estimation of ATT

	Estimate	S.E.	t-val	p-val	CI Lower	CI Upper
(Intercept)	-3.29948	0.07041	-46.86	<2e-16	-3.440	-3.159
treat	0.18083	0.03148	5.74	1.75e-07	0.118	0.244
SIX_MIN	0.00064	0.00024	2.70	0.008	0.000	0.001
VO2max	0.00001	0.00002	0.54	0.591	-0.000	0.000
cUHDRS_Y1	0.26367	0.00396	66.64	<2e-16	0.256	0.272

Table 3: Estimation of ATT

In *table* 3 we report the estimation of ATT (treatment coefficient), its standard error, the associated $p - value$, and the 95 percent confidence intervals. According to *table* 3, the treatment effect is estimated to be equal to 0.181, and the associated $p - value$ is $1.75e - 07 \, (< 0.05)$, which means that the effect of the treatment is considered statistically significant. This is also evident from the largely positive confidence interval. The estimation treatment effect on treated population (ATT) indicates that, *physical activity* leads to a reduction of $cUHDRS$ value, with the amount of decrease on individuals who do not receive the treatment to be on average 0.181 lower compared to individuals who do receive the treatment — thus individuals on treatment group have on average $cUHDRS$ value higher by 0.181, a year after the initial measurement.

The code for the above analysis is provided in 8.4. Outcome and sesitivity analysis parts, performed under the $OVtool \, package$, which is currently under review (*Pane, J., Griffin, B. A.,*

Burgette, L., & McCaffrey, D. Assessing Sensitivity to Omitted Variables: A Tutorial for the OVtool Package. R package. RAND Tool. RAND Corporation. Under review.).

6.6 Assess Sensitivity of the Results to Unobserved Confounding

Sensitivity analysis is performed to assess the robustness of the treatment effect estimation in the presence of potential unobserved confounders (a scenario which is often present). This is performed through a visual tool produced by $OVtool$ package in R.

The output is a contour plot, which represent the treatment effect levels, and the effect of a potential unobserved confounder is considered. 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 group —, 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 (lower-right corner) from the original analysis, which is an indicator of the significance of the effect. The solid black lines of the contour show how the size of the treatment effect changes as a function of the unobserved confounder's relationship with treatment assignment and the outcome, while 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. Considering a fixed level of association of a potential unobserved confounder with outcome and treatment assignment (through correlation and SMD respectively) gives a measure of the shift in treatment effect value, and its significance.
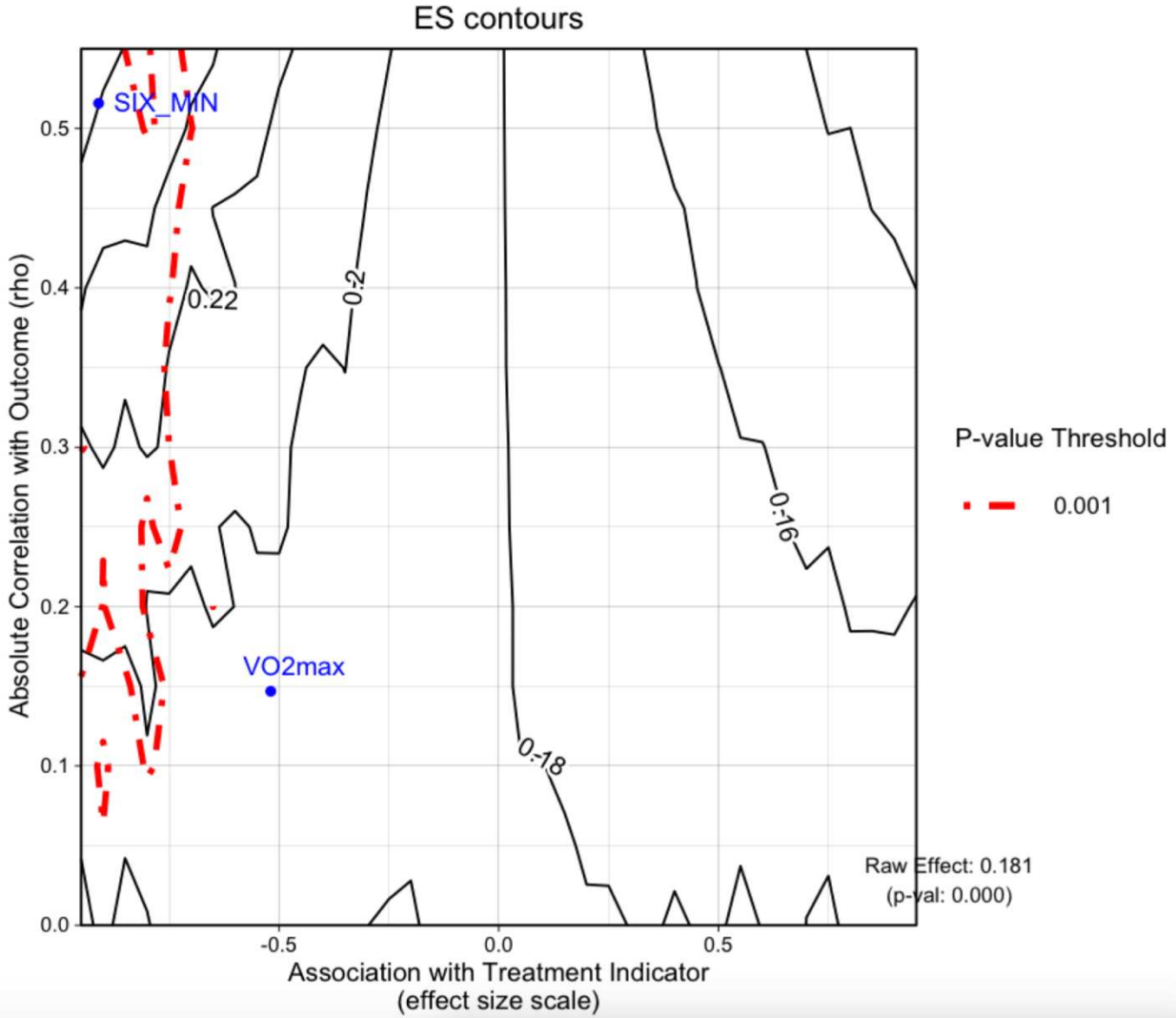
Figure 4: x−axis indicated the SMD of potential confounders and y−axis the correlation with the outcome. of the contour stand for the size of the treatment effect, while the red ones represent the p-value cut-offs. In our case, the observed confounders are V O2max (SMD −0.52, correlation with the outcome 0.15) and SIX MIN (SMD −0.91, correlation with the outcome 0.52)

Figure 4 depicts the output plot of sensitivity analysis for our data. The observed confounders considered for the estimation of treatment effect are $6'\ Walk$, $VO2max$ and $cUHDRS\_Y1$ at the baseline, with SMD values −0.91, −0.52 and 0.15, and association with the outcome (correlation) 0.52, −0.15 and 0.99, respectively (we cannot see the confounder $cUHRDS\_Y1$ in the plot, because of the extreme association with the outcome — correlation almost 1). The summary output of the package is as follows:

- "The sign of the estimated effect is expected to remain consistent when simulated unobserved confounders have the same strength of association with the treatment indicator and outcome

that are seen in the observed confounders. In the most extreme observed case, the estimated effect size is reduced by 7 percent."

- "Statistical significance at the 0.05 level is expected to be robust to unobserved confounders with strengths of associations with the treatment indicator and outcome that are seen in the observed confounders. In the most extreme observed case, the p-value would be expected to increase from 0.000 to 0.002."

Clearly our findings in this case study are extremely robust, most likely due to the inclusion of the baseline outcome in our models. The baseline outcome is always going to be one of the most highly correlated variables with the outcome and thus it explains a lot of the confounding bais in our analysis. We would have to miss something larger than this to wipe away our findings and this is highly unlikely to exist. Consider the *CAP* covariate which is meaningful measure of HD onset in individuals with the HD gene (description in 9) as a candidate unobsered confounder we did not get to include, with *SMD* -0.06 and absolute association with the outcome (correlation) 0.57, the graph above showcases how including this type of confounder would not change our findings in any notable way (again likely because we have the baseline outcome in our list of confounders).

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 & Little, 2017; Olmos & 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, propensity score 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 dealing with 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 scientific advice to choose the covariates we are interested to control for in the PACE-HD dataset, however, there is some literature covering the topic of variable selection for PS models (Brookhart et al., 2006; Hirano & Imbens, 2001).

It is crucial to investigate the data *a priori* for potential overlapping issues 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 access the generalization abilities of the outcome results.

Funding Support

Appendix A. R-Code

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

No technical code required for step 1.

A.2 Assess Sample for any obvious Overlap Concerns and Adjust Sample size as Needed

The code below describes the procedure to compute the mean, standard deviation, minimum and maximum per variable, for each group separately. The treatment status (variable $"treat"$) has entry 0 for the control group, and 1 for the treatment group. Thus, in the beginning, we create vectors with the positions of control and treatment individuals, then computing the desired statistic for each group separately, and finally round them at 2 decimals.

```
control_group <- which(data_full[,"treat"]==0)
treatment_group <- which(data_full[,"treat"]==1)

table_overlap <- cbind(colMeans(data_full[control_group,-1], na.rm=TRUE),
        apply(data_full[control_group,-1], 2, sd, na.rm = TRUE),
        apply(data_full[control_group,-1], 2, min, na.rm=TRUE),
        apply(data_full[control_group,-1], 2, max, na.rm=TRUE),
        colMeans(data_full[treatment_group,-1], na.rm=TRUE),
        apply(data_full[treatment_group,-1], 2, sd, na.rm = TRUE),
        apply(data_full[treatment_group,-1], 2, min, na.rm=TRUE),
        apply(data_full[treatment_group,-1], 2, max, na.rm=TRUE))

table_overlap <- round(table_overlap[,c("SIX_MIN", "VO2max", "cUHDRS_Y1")],2)
```

We produce density plots for each covariate, using $ggplot2$ (install and load the relevant package) to visualize the overlapping of the distributions of the two groups. In each case, we define a $data.frame$ including only two columns (treatment status and the desired covariate), and rename the entries $0/1$ in treatment status to $control/treatment$. In $6'\ Walk$, we add a vertical line ($geom\_vline()$) at the end of treatment group distribution, to underlie the lack of overlapping beyond this point.

```r
library(ggplot2)
##6' Walk
gg_data_6_min <- data_full[,c("treat", "SIX_MIN")]
gg_data_6_min[which(gg_data_6_min[,"treat"]==0),"treat"] <- "Control"
gg_data_6_min[which(gg_data_6_min[,"treat"]==1),"treat"] <- "Treatment"  gg_data_6_min <- data.fram
e(treat=as.factor(gg_data_6_min[,"treat"]),
six_min=as.numeric(gg_data_6_min[,"SIX_MIN"]))

#density plot
p_dens <- ggplot(gg_data_6_min, aes(x=six_min, fill=treat)) +
geom_density(alpha=0.4) + geom_vline(xintercept = 770) +
geom_vline(xintercept = range(gg_data_6_min[gg_data_6_min$treat==
"Treatment", ]$six_min), linetype="dashed", colour="#00BFC4") +
geom_vline(xintercept = range(gg_data_6_min[gg_data_6_min$treat==
"Control", ]$six_min), linetype="dashed", colour="#F8766D") +
labs(title="Distributions of control and treatment group",
y="Rel. Frequency",x="6' Walk",fill="Treatment Group")

p_dens

##VO2max
gg_data_vo2max <- data_full[,c("treat", "VO2max")]
gg_data_vo2max[which(gg_data_vo2max[,"treat"]==0),"treat"] <- "Control"
gg_data_vo2max[which(gg_data_vo2max[,"treat"]==1),"treat"] <- "Treatment"
gg_data_vo2max <- data.frame(treat=as.factor(gg_data_vo2max[,"treat"]),
vo2max=as.numeric(gg_data_vo2max[,"VO2max"]))

#density plot
p_dens <- ggplot(gg_data_vo2max, aes(x=vo2max, fill=treat)) +
geom_density(alpha=0.4) +
geom_vline(xintercept = range(gg_data_vo2max[gg_data_vo2max$treat==
"Treatment", ]$vo2max), linetype="dashed", colour="#00BFC4") +
geom_vline(xintercept = range(gg_data_vo2max[gg_data_vo2max$treat==
"Control", ]$vo2max), linetype="dashed", colour="#F8766D") +
labs(title="Distributions of control and treatment group",
y="Rel. Frequency",x="VO2max",fill="Treatment Group")

p_dens

##cUHDRS
gg_data_cuhdrs <- data_full[,c("treat", "cUHDRS_Y1")]
gg_data_cuhdrs[which(gg_data_cuhdrs[,"treat"]==0),"treat"] <- "Control"
gg_data_cuhdrs[which(gg_data_cuhdrs[,"treat"]==1),"treat"] <- "Treatment"
gg_data_cuhdrs <- data.frame(treat=as.factor(gg_data_cuhdrs[,"treat"]),
cuhdrs=as.numeric(gg_data_cuhdrs[,"cUHDRS_Y1"]))

#density plot
p_dens <- ggplot(gg_data_cuhdrs, aes(x=cuhdrs, fill=treat)) +
geom_density(alpha=0.4) +
geom_vline(xintercept = range(gg_data_cuhdrs[gg_data_cuhdrs$treat==
"Treatment", ]$cuhdrs), linetype="dashed", colour="#00BFC4") +
```

```
geom_vline(xintercept = range(gg_data_cuhdrs[gg_data_cuhdrs$treat==
"Control", ]$cuhdrs), linetype="dashed", colour="#F8766D") +
labs(title="Distributions of control and treatment group",
y="Rel. Frequency",x="cUHDRS",fill="Treatment Group")

p_dens
```

Since in *figure* 3 it is apparent that there is an overlapping problem on the right tail of $6'\ Walk$, we will check for outliers on the right tail of the control group. The code below describes how to find the individuals in the control group with entries over the maximum value of the treatment group.

```
treatment_6_min_max <- max(data_full[treatment_group,"SIX_MIN"])
outliers_obs <- which(data_full[,"SIX_MIN"]>750)

data_full <- data_full[-outliers_obs,]
```

A.3 Estimation of Propensity Scores or Balancing Weights, ideally using Multiple Methods

In this section, we estimate PS for LR, GBM and $CBPS$, and then use the formula described in (5) to obtain the balancing weights, while $EntBal$ estimates directly the balancing weights.

To obtain PS estimation using LR, we consider treatment status ($treat$) as a response, and the rest covariates ($6'\ Walk$, $VO2max$ and $cUHDRS$) as predictors — the same formula is used in every method. As $family$ is used $binomial$ since the nature of treatment status is $0/1$. The fitted values ($fitted()$) of the model are the desired PS, which are transformed with formula (5) to obtain the ATT balancing weights.

```
##LR
set.seed(5)
ps_data <- data.frame(data_full)
##logistic
ps.logit <- glm(treat ~ SIX_MIN + VO2max + cUHDRS_Y1,
        data = ps_data, family = binomial)
PScores <- fitted(ps.logit)
logistic_weights <- with(ps_data,ifelse(treat==1,1,PScores/(1-PScores)))
```

To obtain PS estimation using GBM, we utilize $twang$ package (install and load). The maximum number of trees ($n.trees$), interaction degree ($interaction.depth$), shrinkage level ($shrinkage$) are set to the default values, will we utilize KS_{max} and ES_{mean} as stopping methods (stop.method) — also ES_{max} and KS_{mean} exist. The estimand we wish to estimate is ATT ($estimand$), and finally, we transform the PS according to (5) to derive the ATT balancing weights.

```
library(twang)
##gbm
ps.gbm <- ps(treat ~ SIX_MIN + VO2max + cUHDRS_Y1, data = ps_data,
      n.trees=5000, interaction.depth=2, shrinkage=0.01,
      stop.method=c("es.mean","ks.max"),
      estimand = "ATT", verbose=FALSE)
```

```
PScores <- ps.gbm$ps$es.mean.ATT
gbm_es_weights <- with(ps_data,ifelse(treat==1,1,PScores/(1-PScores)))
PScores <- ps.gbm$ps$ks.max.ATT
gbm_ks_weights <- with(ps_data,ifelse(treat==1,1,PScores/(1-PScores)))
```

CBPS package is used to obtain *CBPS* estimations of PS. *ATT* is set to 1, since we are interested in ATT (0 is used ATE), and the maximum number of iteration for the estimation of the parameters of the model (*iterations*) is set to 1000 (default value). Since *CBPS* returns estimations of PS as fitted values, we transform PS to ATT according to formula (5).

```
library(CBPS)
##cbps
cbps_fit <- CBPS(treat ~ SIX_MIN + VO2max + cUHDRS_Y1, data=ps_data,
        ATT=1, iterations=1000)
PScores <- cbps_fit$fitted.values
cbps_weights <- with(ps_data,ifelse(treat==1,1,PScores/(1-PScores)))
```

Finally, the *entbal* package (*Vegetabile BG. A Tutorial on Entropy Balancing and Weighted Estimation of Causal Effects: A Guide to the 'entbal' Package in R. R package. RAND Tool. RAND Corporation. Under review.*) is used to obtain balancing weight from *EntBal*. The *estimand* we wish to estimate is ATT, and optimization method ($optim_method$) is set to default. The number of moments ($n_moments$) we wish to match here is 1, but one could demand higher moments to be matched — this will deteriorate the ESS value in such small samples though.

```
library(devtools)
require(devtools)
devtools::install_github("bvegetabile/entbal")
library(entbal)
##EB
ps.entbal <- entbal(treat ~ SIX_MIN + VO2max + cUHDRS_Y1, data = ps_data,
        eb_pars = list(exp_type = "binary",
        estimand = "ATT", n_moments = 1, max_iters = 1000,
        verbose = FALSE, optim_method = "L-BFGS-B"))
entbal_weights_1 <- ps.entbal$wts
```

A.4 Assess balance and Effective Sample Size for all methods and choose the best one for Outcome Analysis

The SMD and KS statistic as described in *section* 4, are computed in function $bal.stat$, included in the *twang* package. The function takes as input the data, the names of the variables (*vars*) we with to compute the balance measures, the treatment variable ($treat.var$), the balancing weights ($w.all$), the sample weights (if any) (*sampw*), a verbose for when there are more than two treatment groups (*multinom*) and the estimand of the treatment effect we wish to estimate (*estimand*).

```
rep_1_weights <- rep(1, dim(data_full)[1])
variables_names <- c("SIX_MIN", "VO2max", "cUHDRS_Y1")

unweighted_smd <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=rep_1_weights, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
```

```
multinom=FALSE)$results$std.eff.sz
logistic_smd <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=logistic_weights, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$std.eff.sz
cbps_smd <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=cbps_weights, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$std.eff.sz
entbal_smd_1 <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=entbal_weights_1, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$std.eff.sz
gbm_es_smd <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=gbm_es_weights, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$std.eff.sz
gbm_ks_smd <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=gbm_ks_weights, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$std.eff.sz

unweighted_ks <- bal.stat(ps_data, vars=variables_names,
treat.var="treat", w.all=rep_1_weights, sampw=rep_1_weights, get.ks=TRUE,
estimand="ATT", multinom=FALSE)$results$ks
logistic_ks <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=logistic_weights, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$ks
cbps_ks <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=cbps_weights, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$ks
entbal_ks_1 <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=entbal_weights_1, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$ks
gbm_es_ks <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=gbm_es_weights, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$ks
gbm_ks_ks <- bal.stat(ps_data, vars=variables_names, treat.var="treat",
w.all=gbm_ks_weights, sampw=rep_1_weights, get.ks=TRUE, estimand="ATT",
multinom=FALSE)$results$ks
```

Finally, ESS is computed based on the formula in *section* 4. The function $compute_ess$ tales as input a vector of weights, and returns the ESS.

```
compute_ess <- function(w)
{
   ess_final <- (sum(weights)^2)/sum(weights^2)
   return(ess_final)
}
control_group <- which(data_full[,"treat"]==0)

weights <- rep(1, length(control_group))
unweighted_ess <- compute_ess(weights)
weights <- logistic_weights[control_group]
logistic_ess <- compute_ess(weights)
weights <- gbm_es_weights[control_group]
```

```
gbm_es_ess <- compute_ess(weights)
weights <- gbm_ks_weights[control_group]
gbm_ks_ess <- compute_ess(weights)
weights <- cbps_weights[control_group]
cbps_ess <- compute_ess(weights)
weights <- entbal_weights_1[control_group]
entbal_ess_1 <- compute_ess(weights)
weights <- entbal_weights_2[control_group]

ess <- round(c(unweighted_ess,logistic_ess,gbm_es_ess,gbm_ks_ess,
cbps_ess,entbal_ess_1),1)
```

A.5 Model Outcome and Estimate the Causal Treatment Effect

The estimation of ATT and relevant statistics from the outcome covariate ($cUHDRS_{year2}$ as described in *section* 6.5, are computed by function $outcome_model$, included in the *OVtool* package (*Pane, J., Griffin, B. A., Burgette, L., & McCaffrey, D. Assessing Sensitivity to Omitted Variables: A Tutorial for the OVtool Package. R package. RAND Tool. RAND Corporation. Under review.*). The function takes as input the *weights*, the outcome covariate ($cUHDRS\_Y2$), the model covariates ($VO2max$ and $cUHDRS_{year1}$), and the *estimand* of interest (in our case this is *ATT*).

```
devtools::install_github("jpane24/OVtool")
library(OVtool)
results <- outcome_model(ps_object = NULL,
                stop.method = NULL,
                data = data_full,
                weights = entbal_weights_1,
                treatment = "treat",
                outcome = "cUHDRS_Y2",
                model_covariates = c("SIX_MIN", "VO2max", "cUHDRS_Y1"),
                estimand = "ATT")
output_table <- summary(results$mod_results)$coefficients
```

A.6 Assess Sensitivity of the Results to Unobserved Confounding

ov_sim function of *OVtool* package, takes as input the results of $outcome_model$ function, and evaluates the treatment effect (and it's significance), considering an unobserved confounder, on a range of association with treatment (SMD) and with the outcome (correlation). By default the function considers SMD range $(-0.4,0.4)$, and correlation (rho_grid) range $(0,0.4)$.

The *summary.ov* function, produces a summary report of the sensitivity analysis output.

```
ovtool_results <- ov_sim(model_results=results,
                weight_covariates=c("SIX_MIN", "VO2max"),
                s_grid = NULL,
                rho_grid = seq(0, 0.40, by = 0.20),
                n_reps = 50,
                progress = TRUE,
                add=FALSE)
```

plot.ov(ovtool_results, print_graphic = "3", col = "color")

summary.ov(object = ovtool_results, model_results = results)

Appendix B. Candidate Variables

- *BRUNEL lifestyle physical activity questionnaire (BLPAQ)* is used to collect data on planned and unplanned physical activity. The questionnaire includes 9 questions with five possible answers, scored from 1 to 5. The average score from the first six questions is used as a planned physical activity score (PBRFS) and the average score for the last three questions is used as an unplanned physical activity score (UBRFS).

- *International physical activity questionnaire (IPAQ)*[3]. MET is a measure in minutes and represents the amount of the energy expended carrying out physical activity. MET is a continuous variable and is calculated by adding up products of walking, medium activity, and vigorous activity METs. That is, IPAQ $MET = 3.3 \cdot walkingMET + 4 \cdot mediumMET + 8 \cdot vigorousMET$.

- *Body Mass Index* (BMI) is defined as the body mass divided by the square of the body height and is expressed in units of kg/m^2.

- *CAP score* has been widely utilized in HD related studies to model the effect of age and CAG length on various measures of HD disease. We used the definition introduced in to calculate CAP; $CAP = AGE \cdot (CAG - L)/K$, where L and K are constant. L is an estimate of the lower limit of the CAG expansion at which phenotypic expression of the effect of the mutant huntingtin could be observed, and K is normalizing constant. When $L = 30$ and $K = 6.27$, CAP will be equal to 100 at the subject's expected age of onset of motor symptoms. CAP might be considered as a measure of a subject's cumulative exposure to the toxic effects of mutant huntingtin.

- *Anti-depressant and Anti-psychotic use*, using Anatomical Therapeutic Chemical Classification System (ATC) coding groups such that antidepressants begin with the ATC code N06A and Anti-psychotics, N05A.

- *Six minutes' walk* distance is a maximum distance in meters that HD patient can walk in 6 minutes.

- *Composite Unified Huntington Disease Rating Scale (cUHDRS)* is a composite measure that is used to examine the progressive brain atrophy characteristic of HD in the early HD population. cUHDRS is defined as follows;

$$cUHDRS = \left[\frac{TFC - 10.4}{1.9} - \frac{TMS - 29.7}{14.9} + \frac{SDMT - 28.4}{11.3} + \frac{SWR - 66.1}{20.1}\right] + 10$$

- where TFS, TMS, SDMT, and SWR stand for Total functional capacity, Total Motor score, Symbol Digit Modality Test, and Stroop word reading test, respectively.

- *VO2Max* is a gender-specific measurement that calculates the maximum oxygen consumption during exercise with increasing intensity. It is defined as

$$VO2max \quad = \quad 9.39(W) + 7.7(kg) - 5.88(yr) + 136.7 \; ml \cdot kg^{-1}min^{-1}, \; for \; Females$$
$$VO2max \quad = \quad 10.51(W) + 6.35(kg) - 10.49(yr) + 519.3 \; ml \cdot kg^{-1}min^{-1}, \; for \; Males$$

Bibliography

Agresti, A. (2018). *An introduction to categorical data analysis.* John Wiley & Sons.

Ali, M. S., Groenwold, R. H. H., & Klungel, O. H. (2016). Best (but oft-forgotten) practices: propensity score methods in clinical nutrition research. *The American Journal of Clinical Nutrition*, *104*(2), 247–258.

Altman, D. G., & Bland, J. M. (1999). Treatment allocation in controlled trials: why randomise? *Bmj*, *318*(7192), 1209.

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

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

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

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

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149–1156.

Burgette, L. F., McCaffrey, D. F., & Griffin, B. A. (2015). Propensity score estimation with boosted regression. *Propensity Score Analysis: Fundamentals, Developments and Extensions. New York: Guilford Publications, Inc*.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Choi, B. Y., Wang, C.-P., Michalek, J., & Gelfond, J. (2019). Power comparison for propensity score methods. *Computational Statistics*, *34*(2), 743–761.

Cox, D. R., & Cox, D. R. (1958). *Planning of experiments* (Vol. 20). Wiley New York.

Drew, C. J. G., Quinn, L., Hamana, K., Williams-Thomas, R., Marsh, L., Dimitropoulou, P., Playle, R., Griffin, B. A., Kelson, M., Schubert, R., & others. (2019). Physical activity and exercise

outcomes in Huntington disease (PACE-HD): Protocol for a 12-month trial within cohort evaluation of a physical activity intervention in people with Huntington disease. *Physical Therapy*, *99*(9), 1201–1210.

Elze, M. C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G. W., & Pocock, S. J. (2017). Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology*, *69*(3), 345–357.

Fong, C., Ratkovic, M., Imai, K., & Hazlett, C. (2019). Package 'cbps.' *R Package Version, 0.20*.

Franklin, J. M., Rassen, J. A., Ackermann, D., Bartels, D. B., & Schneeweiss, S. (2014). Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine*, *33*(10), 1685–1699.

Gail, M. H., & Green, S. B. (1976). Critical values for the one-sided two-sample Kolmogorov-Smirnov statistic. *Journal of the American Statistical Association*, *71*(355), 757–760.

Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for constructing and assessing propensity scores. *Health Services Research*, *49*(5), 1701–1720.

Greifer, N. (2020). Package 'cobalt.' *R Package Version, 4.0.0*.

Griffin, B. A., Ayer, L., Pane, J., Vegetabile, B., Burgette, L., McCaffrey, D., Coffman, D. L., Cefalu, M., Funk, R., & Godley, M. D. (2020). Expanding outcomes when considering the relative effectiveness of two evidence-based outpatient treatment programs for adolescents. *Journal of Substance Abuse Treatment*, *118*, 108075.

Griffin, B. A., McCaffrey, D. F., Almirall, D., Burgette, L. F., & Setodji, C. M. (2017). Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of Causal Inference*, *5*(2).

Griffin, B. A., Ramchand, R., Almirall, D., Slaughter, M. E., Burgette, L. F., & McCaffery, D. F. (2014). Estimating the causal effects of cumulative treatment episodes for adolescents using marginal structural models and inverse probability of treatment weighting. *Drug and Alcohol Dependence*, *136*, 69–78.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*(1), 25–46.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*, *15*(3), 234.

Hernán, M. Á., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 561–570.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, *2*(3–4), 259–278.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

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

Kang, J. D. Y., Schafer, J. L., & others. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539.

Lee, J., & Little, T. D. (2017). A practical guide to propensity score analysis for applied clinical research. *Behaviour Research and Therapy*, *98*, 76–90.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390–400.

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

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

Mlcoch, T., Hrnciarova, T., Tuzil, J., Zadak, J., Marian, M., & Dolezal, T. (2019). Propensity Score Weighting Using Overlap Weights: A New Method Applied to Regorafenib Clinical Data and a Cost-Effectiveness Analysis. *Value in Health*, *22*(12), 1370–1377.

Nowok, B., Raab, G. M., Dibben, C., & others. (2016). synthpop: Bespoke creation of synthetic data in R. *J Stat Softw*, *74*(11), 1–26.

Nowok, B., Raab, G. M., Snoke, J., Dibben, C., & Nowok, M. B. (2019). *Package 'synthpop.'*

Olmos, A., & Govindasamy, P. (2015). A practical guide for using propensity score weighting in R. *Practical Assessment, Research, and Evaluation*, *20*(1), 13.

Posner, M. A., & Ash, A. S. (2012). Comparing weighting methods in propensity score analysis. *Unpublished Working Paper, Columbia University*.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., & Griffin, B. A. (2017). Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang package. *Santa Monica, CA: RAND Corporation*.

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., Burgette, L., Burgette, M. L., Ridgeway, M., & Burgette, M. (2020). *Package 'twang.'*

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). *Marginal structural models and causal inference in epidemiology*. LWW.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Setodji, C. M., McCaffrey, D. F., Burgette, L. F., Almirall, D., & Griffin, B. A. (2017). The right tool for the job: Choosing between covariate balancing and generalized boosted model propensity scores. *Epidemiology (Cambridge, Mass.)*, *28*(6), 802.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety*, *17*(6), 546–555.

Wright, R. E. (1995). *Logistic regression.*

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., & Stürmer, T. (2014). The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *American Journal of Epidemiology*, *180*(6), 645–655.

Xie, Y., Zhu, Y., Cotton, C. A., & Wu, P. (2019). A model averaging approach for estimating propensity scores by optimizing balance. *Statistical Methods in Medical Research*, *28*(1), 84–101.