frontiers

Propensity score methods in health technology assessment: principles, extended applications, and recent advances

M Sanni Ali [1–3]*, **Daniel Prieto-Alhambra** [2,4], **Luciane Cruz Lopes** [5], **Dandara Ramos** [3], **Nivea Bispo** [3], **Maria Y. Ichihara** [3,4], **Julia M. Pescarini** [3], **Elizabeth Williamson** [1], **Rosemeire L. Fiaccone** [3,6,7], **Mauricio L. Barreto** [3,6*], **Liam Smeeth** [1*]

[1] *Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom.*
[2] *Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), Center for Statistics in Medicine (CSM), University of Oxford, Oxford, United Kingdom.*
[3] *Centre for Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Muniz, Fundação Osvaldo Cruz, Salvador, Brazil.*
[4] *GREMPAL Research Group (Idiap Jordi Gol) and Musculoskeletal Research Unit (Fundació IMIM-Parc Salut Mar), Universitat Autònoma de Barcelona, Barcelona, Spain.*
[5] *University of Sorocaba - UNISO, Sorocaba, São Paulo, Brasil.*
[6] *Institute of Public Health, Federal University of Bahia (UFBA), Salvador, Bahia, Brazil.*
[7] *Department of Statistics, Federal University of Bahia (UFBA), Salvador, Bahia, Brazil.*

Correspondence*:
M Sanni Ali
sanni.ali@lshtm.ac.uk; sanni.ali@ndorms.ox.ac.uk

ABSTRACT

Randomized clinical trials (RCTs) are considered the gold-standard approach to estimate effects of treatment on outcomes. They are also the designs of choice for health technology assessment (HTA). Randomization ensures comparability, in both measured and unmeasured pre-treatment characteristics, of patients assigned to treatment and control or comparator. However, even adequately powered RCTs are not always feasible for reasons such as cost, time, ethical, and practical constraints. RCTs rely on data collected on selected, homogeneous population under highly controlled conditions; hence, they provide evidence on efficacy of interventions rather than on effectiveness. Alternatively, observational studies can can provide evidence on the relative effectiveness or safety of a health technology compared to one or more alternatives when provided under the routine setting of health care practice. In observational studies, however, treatment assignment is a non-random process based on subjects baseline characteristics hence treatment groups may not be comparable in their pre-treatment characteristics. As a result, direct comparison of outcomes between treatment groups lead to biased estimate of treatment

effect. Propensity score methods have been used to achieve comparability of treatment groups in terms of their measured pre-treatment covariates and thereby controlling for confounding bias in estimating treatment effects. Despite the popularity of propensity scores methods and recent important methodological advances, misunderstandings on their applications and limitations are all too common. In this article, we provide a review of the propensity scores methods, extended applications, recent advances, and strengths and limitations.

Keywords: bias, confounding, effectiveness, health technology assessment, propensity score, safety, secondary data, observational study

1 INTRODUCTION

Randomized clinical trials (RCTs) are considered the gold-standard approaches for estimating the "causal" effects of treatments on outcomes (Sibbald and Roland, 1998; Concato et al., 2000) and the design of choice for health technology assessment. In causal inference terminology using Rubin's potential outcomes framework (Rubin, 2005), the effect of a certain treatment (T = 1) versus a control or comparator (T = 0) on an outcome (Y) involves comparison of potential outcomes under treatment (Y_1) and an alternative treatment (Y_0). In RCT, with sufficient numbers of participants and adequate concealment of allocation, randomization ensures that individuals assigned to treatment and control or comparator groups are comparable in all pre-treatment characteristics, both measured and unmeasured (Sibbald and Roland, 1998). The only difference is that one group recieved the treatment (T = 1) and the other recieved no treatment or alternative treatment (T = 0), hence, any difference in outcomes between the two groups can be attributable to the effect of the treatment. In other words, the "causal" effect of treatment in the study population (the average treatment effect, ATE) on outcomes can be estimated by a direct comparison of the expected outcomes between the treatment and the comparator groups (Equation 1) (Concato et al., 2000). However, even adequately powered RCTs may not always be feasible for reasons such as cost, time, ethical, and practical constraints (Sibbald and Roland, 1998). RCTs also rely on data collected on selected, homogeneous population under highly controlled conditions; hence, they provide evidence on efficacy rather than on effectiveness of interventions or treatments (Eichler et al., 2011).

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

With steadily increasing health care costs and the introduction of novel, yet expensive, pharmaceutical products, health technology assessment (HTA) agencies such as the UK National Institute for Health and Care Excellence (NCIE) are inquiring robust methods for evaluation of comparative effectiveness and safety of medications, devices, and diagnostics in routine clinical practice. In contrast to efficacy, comparative effectiveness of an intervention is the extent to which an intervention does more good than harm, when compared to one or more alternative intervention(s) for obtaining the desired results when used under the routine setting of health care practice (Eichler et al., 2011; Schneeweiss et al., 2011). In addition, for medical devices and diagnostics, waiting for evidence from RCTs when the health technology is diffusing in the clinical practice could be costly for the payers, inefficient from policy perspective, and methodologically questionable (Tarricone et al., 2016). On the other hand, regulators' and HTA agencies' perception of the value of real world data in enriching evidence on the effectiveness of health technologies has been steadily increasing (Yuan et al., 2018).

The effect of a particular health technology such as a treatment on a certain outcome could also be investigated using non-randomized studies (i.e., observational or quasi-experimental) using routinely collected clinical data (Schneeweiss et al., 2011; Bärnighausen et al., 2017). In observational studies, however, treatment selection is influenced by patient, physician, and, to a certain extent, health system characteristics. Treated and untreated groups differ not only in receiving the treatment but also in other pre-treatment characteristics, leading to non-comparability or non-exchangeability, a phenomenon known as confounding bias (Greenland and Morgenstern, 2001). This means that differences in outcomes between the treated and untreated groups could be explained by either the treatment, or other pre-treatment variables, or both. In other words, direct comparison of outcomes between treated and untreated groups leads to biased estimate of the treatment effect. Hence, any systematic difference in pre-treatment characteristics should be accounted for either by design or analysis (Rubin, 1997). Over the years, several methodologies have been developed to control for confounding bias in observational studies (Figure 1); the propensity score methods (Rosenbaum and Rubin, 1983) are among the popular approaches in pharmacoepidemiology and health technology evaluations (Ali et al., 2015).

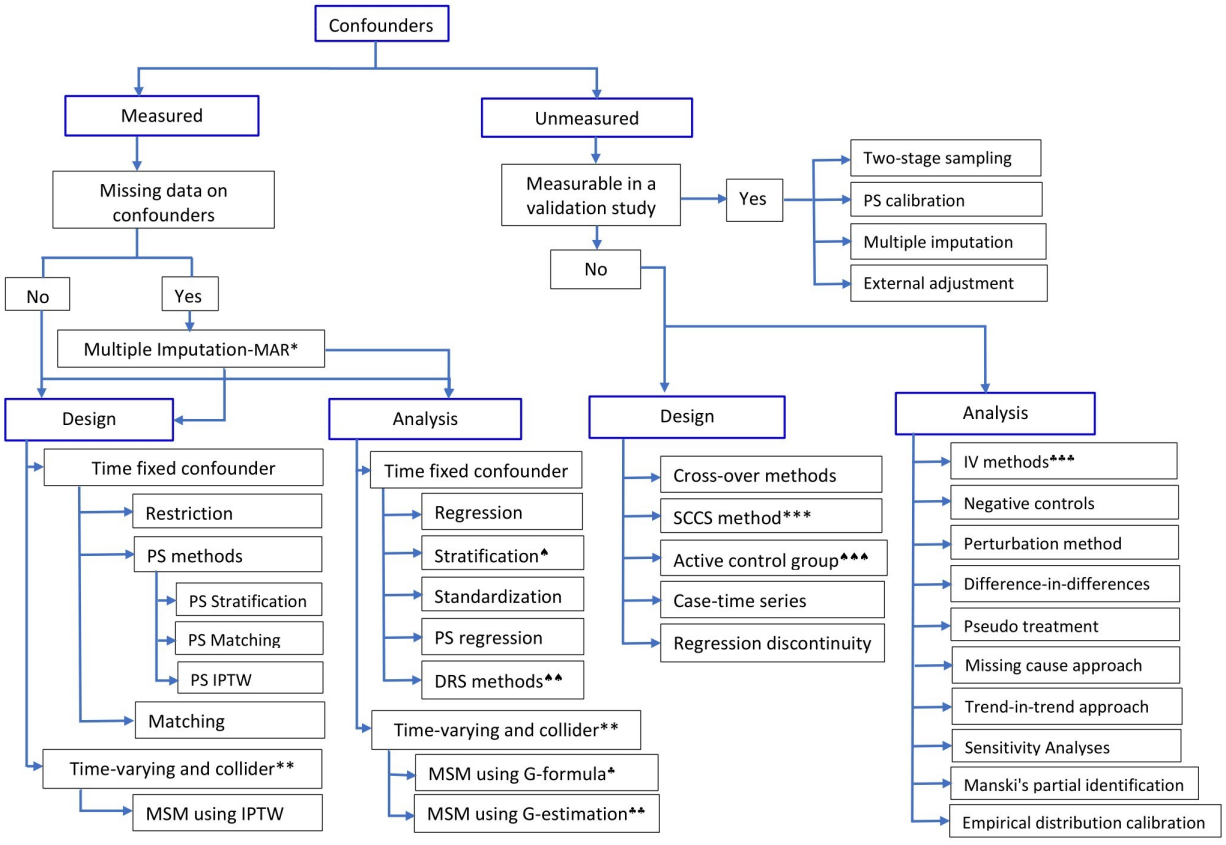

Figure 1. Methods to control for confounding in observational studies.*Multiple imputation is valid under the assumption of Missing at Random (MAR);**If time-varying confounder is affected by previous treatment, all PS-based methods except MSM using IPTW will give biased estimate;*** Self-controlled case-series design; ♠stratification using effect modifier and adjustment within the strata to account for other covariates; ♠♠Disease risk score (prognostic score) method; ♠♠ restriction or choosing an active comparison group vs non-user group; ♣ G-formula and ♣♣G-estimation of structural nested models; ♣♣♣Instrumental variable methods. (Adapted in part from Schneeweiss (Schneeweiss, 2006), Uddin et al. (Uddin et al., 2016), and Zhang et al. (Zhang et al., 2018)

Propensity score methods were introduced by Rosenbaum and Rubin in 1983 (Rosenbaum and Rubin, 1983) and their use to control for confounding has been increasing in the previous decade. Propensity score is a scalar summary of all measured potential confounders; stated formally, the propensity score e(x) is the conditional probability of receiving a certain treatment, versus a comparator/no treatment, given the measured pre-treatment chracteristics (Rosenbaum and Rubin, 1983), x, denoted as

$$e(x) = pr(z = 1|x), \tag{2}$$

where z = 1 for subjects in the treatment group and z = 0 for subjects in the comparison group (Rosenbaum and Rubin, 1983, 1984). Treated and untreated subjects with similar propensity scores have, on average, similar pre-treatment characteristics, a situation similar to RCTs. However, this conditional comparability, given the propensity score, of the treatment groups is limited only to measured pre-treatment characteristics included in the propensity score model and not unmeasured ones (Rosenbaum and Rubin, 1983). Hence, balancing these pre-treatment potential confounders through propensity scores enables researchers to obtain "quasi-randomization" of treatment groups to minimize confounding and to get a better estimate of the effect of treatment. Implicitly, researchers assume "strongly ignorable treatment assignment" (SITA) given the measured covariates, this comprises "unconfoundedness" and "positivity" (Rosenbaum and Rubin, 1983). Unconfoundedness implies that all relevant pre-treatment characteristics are measured and included in the propensity score model hence given these measured covariates are included in the propensity score there is no unmeasured confounding. Positivity, on the other hand, implies that every individual has a non-zero (positive) probability of receiving all values of the treatment variable: $0 < P(z = 1| X) < 1$ for all values of z (Rosenbaum and Rubin, 1983).

In the last decade, the propensity score methods have been increasing popular among clinical researchers, their use in pharmacoepidemiology and health technology assessments has been ubiquitous, and have undergone substantial methodological advances. On the other hand, confusions and misunderstandings on what a propensity score methods can and cannot do as well errors in the design, analysis, interpretation, and reporting of propensity score based analyses are unfortunately all too common (Ali et al., 2015). With increasing availability of routinely collected electronic medical records for evaluation of effects (both comparative effectiveness and safety) of health technologies, and relatively rapid development of the methods, an up-to date review of the methods and their characteristics is necessary. In this article, we aim to introduce propensity score methods with an emphasis on important aspects of the methods; describe their extended applications and recent developments; and discuss their strengths and limitations.

The manuscript, including the introduction, is organized in to eight sections: section one introduces RCTs, observational studies and propensity score in relation to health technology assessment; section two discusses variable selection and propensity score estimation approaches; section three describes covariate balance assessment in propensity score methods; section four summarizes the different types of propensity score methods; section five describes extended applications of propensity scores; section six summarizes strengths and limitations of the propensity score methodology; section seven highlights on reporting of propensity score based analysis; and section eight concludes the discussion.

2 VARIABLE SELECTION AND PROPENSITY SCORE ESTIMATION

Observational studies using administrative or clinical databases often involve high-dimensionality with respect to the number of covariates available for analysis including socio-economic characteristics, demographics, co-morbidities, co-medications, health system characteristics, among others. The inclusion

of large number of variables in conventional regression models, particularly in non-linear models such as logistic and Cox regression models, requires sufficient number of events (approximately 10 events to account for every confounder included in the regression model) (Peduzzi et al., 1995, 1996; Cepeda et al., 2003). For example, to adjust for 5 confounders using logistic regression model, one would need to have 5*10 = 50 outcome events. However, many practical settings in pharmacoepidemiology and other health technology assessments involve relatively few or rare outcome events, hence, confounding adjustment using regression methods requires selection of a limited number of covariates to avoid problems such as over-fitting (Peduzzi et al., 1995). Alternatively, the use of propensity score methods to summarize a large pool of covariates in to a single score, the propensity score, avoids over-fitting and collinearity issues in estimating treatment effects (Cepeda et al., 2003). When the number of covariates available in the study dataset is relatively small, it is a common practice to include all the covariates in the propensity score model; however, covariate selection might be required when researchers are presented with very large number of covariates (several hundreds) and limited number of events (Schneeweiss et al.,2009).

Covariates selection in propensity score is often based on prior subject-matter knowledge on the relations underlying the covariates in the data at hand, statistical tests on the association between the covariates and the outcome (using p-values or change in effect estimates) (Brookhart et al., 2006; Patrick et al., 2011; Ali et al., 2015; Adelson et al., 2017), strength of associations with treatment and/or outcome (Patrick et al., 2011; Ali et al., 2015; Adelson et al., 2017), and machine learning methods (McCaffrey et al., 2004). Each approach has its own strengths and limitations, however, emphasis should be given to achieve balance on important prognostic pre-treatment characteristics (Rosenbaum and Rubin, 1983) and not to improve model fit or to predict treatment as well as possible. Hence, the use of p-values, goodness-of-fit tests, and model discrimination tests such as c-statistics should be avoided (Weitzen et al., 2005; Patrick et al., 2011; Westreich et al., 2011). The iterative approach of model fitting, by including interactions and square terms of the covariates, and subsequent balance assessment, which was recommended in the seminal paper by Rubin and Rosenbaum (Rosenbaum and Rubin, 1983), is still a more robust approach. This approach helps to achieve the goal of propensity score modelling, ”improving balance” of potential confounders between treatment groups so that the groups are comparable or exchangeable conditional on the propensity score.

One of the greatest strengths of propensity score approaches is the separation of design from analysis, i.e., propensity score methods purposefully disregard outcome information at this stage of the design (Rubin, 2004b; Leacy and Stuart, 2014). That would also mean, as in the classical implementation of the methods, association between the covariates and the outcome in the study data is not assessed for selection of covariates to construct the propensity score model. However, this appraoch is not without disadvantages: failure to include colliders (common effects of treatment and outcome related covariates) and exclude strong instruments (variables strongly related to treatment but independent of both confounders and outcome) can lead to increased bias (Pearl, 2011, 2012; Myers et al., 2011a,b; Ali et al.,2016).

It is important to emphasize that, similar to conventional regression modelling, intermediates (variables on the causal pathway between treatment and outcome) and colliders should not be included in the propensity score model (Greenland and Morgenstern, 2001) since including these variables will tend to increase (rather than reduce) bias. In addition, strong instruments should also be excluded, particularly when strong unmeasured confounding is a concern thereby avoiding any amplification of the residual bias (Pearl, 2011, 2012; Myers et al., 2011a,b; Ali et al., 2016). However, it is unusual to encounter such a scenario; the use of propensity score method is only meaningful when the assumption of strongly ignorable treatment assignment is met (i.e., there is no unmeasured confounding given the measured covariates and also there is positivity) (Rosenbaum and Rubin, 1983). Bias amplification should be considered a secondary

concern compared to residual confounding by unmeasured characteristics hence researchers should err on the side of inclusion rather than exclusion of potential confounders (Myers et al., 2011b; Ali et al., 2017c). Alternatively, when a strong instrument - essentially a proxy measure of difference in treatment - is identified that is independent of confounders and outcome, instrumental variable analysis can be a powerful tool to account for any unmeasured confounding (Angrist et al., 1996).

A common question asked by clinical researchers who have not used propensity score methods is why do we estimate the probability that an individual receives a certain treatment versus a comparator while we certainly know from the data whether that individual has received the treatment. A brief answer to this question is as follows: propensity score exists both in RCTs and in observational studies. In RCTs, the true propensity score is known and is defined by the study design or treatment allocation mechanism, i.e., randomization. For example, in a simple two-arm RCT in which individuals are assigned to a treatment or a comparison group by a flip of a fair coin (assuming equal sample sizes in both treatment groups), the propensity score for every individual is the probability of being assigned to the treatment group vs. the comparator group, which is equal to 0.5, apart from chance variations. In contrast, in observational studies, the true propensity score for individuals is unknown and is dependant on several pre-treatment characteristics, both clinical and non-clinical, under consideration by the physician. As a result, it should be - and can often be - estimated using the study data (D'Agostino Jr, 2007; Joffe and Rosenbaum, 1999; Rubin, 2004b; Ali et al., 2016). Estimation of the propensity score is needed to create a "quasi-randomized experiment" by using the individuals's probability of receiving the treatment as a summary score of all measured pre-treatment covariates. It enables appropriate adjustment for measured potential confounders to estimate the effect of the treatment. This explains one of the key properties of the propensity score method: if we find two individuals with the same propensity score, one in the treated group and one in the untreated group, we can assume that these two individuals are more or less "randomly assigned" to one of the treatment groups in the sense of being equally likely to be treated or not, with respect to measured pre-treatment characteristics Ali et al. (2015, 2016).

In practice, the propensity score is often estimated using a logistic regression model, in which treatment status is regressed on measured baseline characteristics (Austin, 2008a; Ali et al., 2015). The estimated propensity score is the predicted probability of treatment derived from the fitted regression model. Logistic regression has several advantages: it is a familiar and well-understood statistical tool for researchers as well as easy to implement using standard statistical software packages (Setoguchi et al., 2008; Westreich et al., 2010). However, logistic regression is not the only approach; other methods have also been used including recursive partitioning (D'Agostino Jr, 2007) and several machine learning methods such as classification and regression trees (CARTs), neural networks, random forests, among others (Setoguchi et al., 2008; Lee et al., 2010, 2011; Westreich et al., 2010). Comparative simulation studies favour the use of machine learning methods over logistic regression when there is moderate or high non-linearity (interactions between baseline covariates) and non-additivity (square or cubic terms) in the propensity score models. This could be explained by the fact that machine learning methods include interaction and square terms by default (Setoguchi et al., 2008), compared to logistic regression where the researcher should "manually" include interaction and square terms. When important interaction and square terms are included, the performance of logistic regression is as good as other machine learning methods (Ali et al., 2017b).

3 COVARIATE BALANCE ASSESSMENT

The aim of propensity score methods is to balance covariates between treatment groups hence control for measured confounding (Rosenbaum and Rubin, 1983). Therefore, the quality of propensity score model should be assessed primarily on the covariate balance achieved. It should not be evaluated based on how well the propensity score model discriminates between treated and untreated individuals, i.e., whether the treatment process is correctly modeled (Rubin, 2004b; Westreich et al., 2011; Ali et al., 2015, 2016) or whether the eventual treatment effect estimates are larger or smaller than expected (Rosenbaum and Rubin, 1984; Hansen, 2004). Hence, propensity score model fitting can be considered as an iterative step where the propensity score model is updated by adding different covariates, interactions between covariates, or higher-order terms of continuous covariates until an acceptable balance on important confounders is achieved (Rosenbaum and Rubin, 1984). It is also important to underline that variable selection and covariate balance are inseparably linked; however, covariate balance is often checked on a pre-selected list of pre-treatment covariates (Ali et al., 2015). On the other hand, there are propensity score modelling techniques that take in to account covariate balance while selecting covariates for the propensity score model (Imai and Ratkovic, 2014; Austin, 2019).

It is helpful to start propensity score analysis by examining the distribution of propensity scores using histograms or density plots. This facilitates subjective judgment on whether there is sufficient overlap, also called "the common support", between propensity score distributions of treated and untreated groups (Dehejia and Wahba, 2002). However, such plots should not considered as proper measures of covariate balance;they can guide the choice of matching algorithms in propensity score matching and the number of strata in propensity score stratification (Ali et al., 2015, 2016). For example, when there is very little overlap in the propensity score distributions, matching treated and untreated subjects with replacement, with or without caliper, can be a better option because it will be challenging to find sufficient number of untreated subjects for the treated subjects (Ali et al., 2016). Inadequate overlap in the propensity score distributions, which can be quantified using overlapping coefficient Ali et al. (2014), should also warn researchers that the dataset, no matter how large, could not support any causal conclusion about the effect of the treatment on the outcome of interest (Rubin, 1997; Ali et al., 2016).

To assess covariate-specific balance, several metrics have been proposed in the literature (Belitser et al., 2011; Austin, 2009; Ali et al., 2014). Each balance metric has its own advantages and limitations; the absolute standardized difference in means or proportions (ASMD) (Austin, 2009) is more robust in terms of covariate distributions and sample size requirements compared to other balance metrics, such as overlapping coefficients (Ali et al., 2014, 2015, 2016). The ASMD is a well-understood and easy to calculate statistical tool hence it is recommended for checking and reporting covariate balances in propensity score methods (Austin, 2009; Belitser et al., 2011; Ali et al., 2014, 2015, 2016). The ASMD is calculated for each covariate and can be averaged to compute an overall covariate balance and to compare propensity score models (Belitser et al., 2011; Ali et al., 2014). The covariate-specific ASMD is useful to identify the variable that is still imbalanced and to modify the propensity score model with squares and interaction terms of the variable to improve its balance. Although there is no universal threshold below which the level of imbalance is always acceptable (Imai and Van Dyk, 2004; Ali et al., 2016), the use of arbitrary cut-offs for balance diagnostics (e.g., <10% for the ASMD) is common in the medical literature (Ali et al., 2015, 2016). Balance is not only a property of the sample means of a covariate but of the overall distribution, hence, higher-order sample moments of the distribution such as variance should also be evaluated (Rosenbaum and Rubin, 1985; Rubin, 2001; Ho et al., 2007; Austin, 2009; Linden and Samuels, 2013). Rubin (Rubin, 2001) proposed the ratio of variances of treated and untreated groups as a balance measure; a variance ratio

of 1 in the matched sample indicates a good matching, and a variance ratio below 2 is generally considered acceptable (Rubin, 2001; Linden and Samuels, 2013).

In addition to numerical quantification of the covariate balance achieved by the specified propensity score model, graphical methods such as quintile-quintile plots; side-by-side (weighted) box plots; plots of absolute standardized differences of means (ASMD); and empirical density plots for comparing the distribution of continuous baseline covariates provide a simplified overview of whether balance on individual pre-treatment covariate has improved (Rosenbaum and Rubin, 1983; Ali et al., 2016).

4 PROPENSITY SCORE METHODS

Once the propensity score has been estimated, researchers have several options regarding how to use the propensity score in the analyses, including matching, stratification (also called sub-classification), covariate adjustment using the propensity score, inverse probability of treatment weighting, and combination of these methods (Rosenbaum and Rubin, 1983, 1984; Rubin and Thomas, 2000; Hirano and Imbens, 2001; Johnson et al., 2018). Each method has its own advantages and disadvantages; the choice of the propensity score method is in part determined by the inferential goal of the research (i.e., the type of treatment effect estimand: the average treatment effect in the entire population, ATE, versus the average treatment effect in the treated population, ATT) (Imbens, 2004; Stuart, 2008; Ali et al., 2016). Although it is possible to estimate both ATT and ATE using all of the four propensity score methods, for example, by assigning different weights for the treated and untreated individuals, the default approach in each method might give slightly different estimand. For example, propensity score matching typically estimates the treatment effect in the treated group, ATT (Imbens, 2000). Therefore, to get an estimate of the average treatment effect in the entire population, ATE, one has to use either full matching (Hansen, 2004) or different weighting (Stuart, 2008, 2010; Ali et al., 2015, 2016). The use of a specific propensity score method has also direct implication on covariate balance assessment (Rosenbaum and Rubin, 1983, 1984; Ali et al., 2016) and interpretation of the estimated treatment effect (Stuart, 2008; Ali et al., 2015, 2016).

4.1 Propensity score matching

Propensity score matching, the most common application of propensity score (Ali et al., 2015), entails forming matched sets of treated and untreated subjects who share a similar value of the propensity score (Rosenbaum and Rubin, 1983; Rubin and Thomas, 1996). The matching could be done in many ways: one-to-one or one-to-many (1:n, where n is the number of controls often up to five) matching with or without replacement, stratified matching, and full matching (Hansen, 2004). However, one-to-one caliper matching without replacement is the most common implementation of propensity score matching (Ali et al., 2015, 2016). For detailed discussion on different matching approaches, we refer to the literature (Rosenbaum and Rubin, 1985; Hansen, 2004; Stuart, 2010).

Once a matched sample has been formed, covariate balance can be easily checked between the matched groups using one of the balance metrics, preferably ASMD, and treatment effect can be estimated by directly comparing outcomes between treated and untreated subjects in the matched sample (Rosenbaum and Rubin, 1983; Rubin and Thomas, 1996). With dichotomous or binary outcomes (e.g., the presence of a disease, ”Yes” or ”No”), the effect of the treatment can be estimated as the difference or ratio between the proportion of subjects experiencing the outcome event in each of the two treatment groups (treated vs. untreated) in the matched sample. If the outcome is continuous (e.g., blood pressure measurement or HBA1c level), the effect of the treatment is estimated as the difference between the mean outcome for

treated individuals and the mean outcome for untreated individuals in the matched sample (Rosenbaum and Rubin, 1983).

If matching is done with replacement or in one-to-many matching, weights should be incorporated to account for the multiple use of the same untreated individual to match with several treated individuals or the multiple use of the same treated individual to match with several untreated individuals, respectively (Stuart, 2010). Whether to account for the matched nature of the data in estimating the variance of the treatment effect, hence, the use of paired t-test for continuous outcome or McNemar's test for binary outcome, is an ongoing discussion (Stuart, 2008; Schafer and Kang, 2008; Austin, 2008a, 2011).

The most appealing feature of propensity score matching is that the analysis can partly mimic that of an RCT, meaning that the distribution of measured baseline covariates will be, on average, similar between treatment groups. Hence, direct comparison of outcomes between treated and untreated groups within the propensity score matched sample has the potential to give unbiased estimate of the treatment effect, depending on the extent to which the measured variables have captured the potential confounding factors (Rosenbaum and Rubin, 1983). However, RCT, on average, guarantees balance on both measured an unmeasured confounders whereas propensity score improves balance on measured confounders but those of unmeasured confounders only to the extent that they are related to the measured confounders included in the propensity score model (Rubin, 2004b; Austin, 2011). Other useful features include: separation of the design from analysis via pre-processing of the data to improve covariate balance without using outcome data, thereby a minimum reliance on model specification; relatively easy assessment, visualization, and communication of covariate balance using simple statistics or plots; and qualitative indication of whether the dataset at hand is good enough to address the causal question without relying on untrustworthy "model-dependent" extrapolations (Rubin, 2004b; Ho et al., 2007; Ali et al., 2016).

Recently, the use of propensity score for matching has been criticized on the basis of an argument that propensity score matching approximates complete randomization and not completely blocked randomization, hence, it engages in random pruning or exclusion of individuals during matching. "Unlike completely blocked randomization, random exclusion of individuals in propensity score matching, as in complete randomization, means a decrease in sample size leading to covariate imbalance and more model dependence, so called the "propensity score paradox"" (King and Nielsen, 2016). At first this might seem a valid argument; however, the practical implication of this paradox is very limited, if any (Ali et al., 2017a). This is partly due to the fact that propensity score matching could do better than complete randomization with respect to the balance of measured covariates if variables related to treatment are included in the propensity score model (Joffe and Rosenbaum, 1999). In addition, the use of matching algorithms such as caliper matching or matching with replacement retains the best matches thereby avoiding random prunning or exclusion, and hence the paradox is not a big concern. Furthermore, it is currently a standard practice to check covariate balance in the propensity score matched sample before estimating the treatment effect, further minimizing any risk of exacerbating covariate imbalance (Ali et al., 2015).

Similar to RCTs, when there are residual differences in baseline characteristics between treatment groups in propensity score matched sample, regression adjustment can be used on the matched sample to reduce bias due to residual differences in important prognostic factors (Rubin and Thomas, 2000; Imai and Van Dyk, 2004; Schafer and Kang, 2008). This method has been described as a doubly robust approach, i.e., correct specification of either the matching or the regression adjustment, but not necessarily both, is required to obtain unbiased estimate of the treatment effect (Schafer and Kang, 2008; Nguyen et al., 2017; Funk et al., 2011). Propensity score matching primarily estimates the effect of treatment in the treated individuals (ATT), not the the effect of treatment in the population(treated and untreated individuals,

ATE) (Imbens, 2004; Stuart, 2008), this is because the closest untreated matches are selected for treated individuals, and unmatched untreated subjects are often excluded from the analysis (Stuart, 2008; Ali et al., 2016). It is important to emphasize that exclusion of unmatched subjects from the analysis not only affects the precision of the effect estimate but also could have consequences for the generalizability of the findings, even for the ATT (Lunt, 2013; Ali et al., 2016). For example, exclusion of untreated subjects due to a lack of closer matches could change the estimand from the effect of treatment in the treated individuals (ATT) to the effect of treatment those treated individuals for whom we can find untreated matches(ATT') (Lunt, 2013). However, it is possible to estimate the ATE in the matched sample with modifications of the matching algorithms. For example, full matching, which retains all the subjects in the data for analysis, can estimate either the ATT or the ATE (Hansen, 2004; Stuart, 2010). Although matching, in general, discards some data (i.e., unmatched subjects), it can actually increase the efficiency of the treatment effect estimate (Ho et al., 2007; Ali et al., 2016).

4.2 Propensity score stratification

Propensity score stratification, also called propensity score sub-classification, involves grouping individuals in to strata based on their propensity scores (often five groups using quintiles or ten groups using percentiles). Within these strata, treated and untreated individuals will have a similar distribution of measured covariates; hence, the effect of the treatment can be estimated by direct comparison of outcomes between treated and untreated groups within each strata (Rosenbaum and Rubin, 1984; D'Agostino Jr, 2007; Ali et al., 2016; Adelson et al., 2017). The stratum-specific treatment effects can then be aggregated across sub-classes to obtain an overall measure of the treatment effect (Rosenbaum and Rubin, 1984).

Rosenbaum and Rubin (Rosenbaum and Rubin, 1983, 1984) proposed quintile stratification on the propensity score based on their finding that five equal-size propensity score strata removed over 90% of the bias due to each of the pre-treatment covariates used to construct the propensity score. However, it is recommended that researchers examine the sensitivity of their results to the number of sub-classes by repeating the analysis using different quantiles of the propensity score (Imai and Van Dyk, 2004; Adelson et al., 2017). Similar to matching, residual imbalances after stratification can be accounted for using regression adjustment within each stratum (Rubin, 2001; Rosenbaum and Rubin, 1984). Alternatively, the quintiles and deciles of the propensity score can be used as a categorical variable in a model-based adjustment to estimate treatment effects (Rosenbaum and Rubin, 1984).

Propensity score stratification can estimate either the stratum-specific or overall ATT or ATE depending on how the subclass estimates are weighted. Weighting stratum-specific estimates by the proportion of treated individuals in each stratum provides ATT, whereas weighting by the total number of individuals (treated and untreated) in each stratum yields the ATE (Stuart, 2010). Similarly, pooling stratum-specific variances provides pooled estimates of the variance for the pooled ATT or ATE estimate. Pooling the stratum-specific treatment effect is straightforward when treatment effect is homogeneous among the propensity score strata. When there is heterogeneity of treatment effect among the strata even after automated iterations of the number and boundaries of propensity score strata (Imbens, 2004; Imbens and Rubin, 2015), pooling the stratum-specific treatment effect complicates interpretation of the treatment effect (Ali et al., 2014, 2016). In the presence of effect-measure modification and regardless of the presence of confounding, Mantel-Haenszel methods do not estimate a population parameters (ATE) hence estimating the effect of treatment in the treated (ATT) rather than the whole population (ATE), for example, using propensity score matching is preferable (Stürmer et al., 2006b). Alternatively, one could standardize the stratum-specific

estimates to a specified distribution of propensity scores, for example, to calculate a standardized mortality ratio (AMR) from the stratum-specific estimates (Lunt et al., 2009).

Stratification has several advantages: it is easy to implement; it is straightforward to evaluate and communicate covariate balance, and to interpret particularly to non-technical audiences; it separates the design of the study from the analysis, like propensity score matching, hence less dependant on parametric models (Rosenbaum and Rubin, 1984); it is less sensitive to non-linearities in the relationship between propensity scores and outcomes; and it can accommodate additional model-based adjustments (Rosenbaum and Rubin, 1983, 1984).

4.3 Regression adjustment using propensity score

The propensity score, as a single summary of all covariates included in the propensity score model, can be included as a covariate in a regression model of the treatment, i.e., the outcome variable is regressed on the treatment variable and the estimated propensity score. Although this approach is very easy to implement, it is generally considered to be a sub-optimal application of the propensity score for several reasons: 1) The treatment effect estimation is highly model-dependent because it mingles the study design and data analysis steps, hence, it requires correct specification of the propensity score model (Rubin, 2004b; Johnson et al., 2018). 2) It also makes additional assumptions unique to regression adjustment; the relationship between the estimated propensity score and the outcome must be linear and there should be no interaction between treatment status and the propensity score (Rosenbaum and Rubin, 1983; Austin, 2011; Ali et al., 2016). 3) It enables estimation of the ATE, however, its interpretation is complicated particularly in non-linear models such as logistic regression or Cox regression where the estimand of interest is non-collapsible. Non-collapsibility refers to a phenomenon in which, in the presence of a non-null treatment effect, the marginal (overall) treatment effect estimate is different from the conditional (stratum-specific) treatment effect estimate, even in the absence of confounding (Greenland et al., 1999; Austin, 2008b). In addition, assessment and communication of covariate balance is not straightforward (Ali et al., 2016).

4.4 Inverse probability treatment weighting

Inverse probability weights (IPWs) calculated from propensity score can be used to create an "artificial" population, also called a "pseudo-population," in which treatment is independent of measured pre-treatment characteristics (Robins et al., 2000; Hernán et al., 2000; Cole and Hernán, 2008). Hence, treated individuals will be assigned weights equal to the inverse of their propensity scores (1/PS, as they have received the treatment) and untreated individuals will be assigned weights equal to the inverse of one minus their propensity scores (1/(1 – PS)) (D'Agostino Jr, 2007). A particular diagnostic concern with regard to propensity score weighting is that individuals with extremely large weights may unduly influence results and yield estimates with high variance (Lee et al., 2011). When some individuals have probabilities of receiving the treatment close to 0 or 1, the weights for such individuals become extremely high or extremely low, respectively. Weight stabilization to "normalize" the range of the inverse probabilities is often considered: the "1" in the numerator of the inverse probability weights can be replaced with the proportion of treated individuals and the proportion of untreated individuals for treated and untreated individuals, respectively (Hernán et al., 2000; Ali et al., 2016).

Alternative approaches such as weight trimming and weight truncation have been suggested (Cole and Hernán, 2008; Lee et al., 2011). Weight trimming involves removing individuals in the tails of the propensity score distributions using percentile cut-points (Cole and Hernán, 2008; Lee et al., 2011), i.e., individuals who have extreme values of the propensity score - both very high and very low are excluded. On

the other hand, weight truncation involves setting a maximum allowable weight, w_{ma}, such that individuals with a weight greater than w_{ma} will be assigned w_{ma} instead of their actual weights. Both approaches may help stabilize weights, reduce the impact of extreme observations, and can improve the accuracy and precision of parameter estimates; however, both involve bias-variance trade-offs (Lee et al., 2011). For example, trimming the tails excludes some individuals with extreme values hence change the population which might introduce bias depending on the cut-off (Cole and Hernán, 2008). Recently, Li et al (Li et al., 2018) suggested a different set of weights called overlapping weights which weight each individual proportional to its probability of receiving to the alternative treatment. Unlike standard IPWs, the overlap weights are bounded between 0 and 1 and thus are less sensitive to extreme weights. It also means that there is no need for arbitrary choice of a cut-off for inclusion in the analysis as well as exclusion of individuals, unlike weight trimming (Li et al., 2018).

In the weighted population, also called ”psudo-population”, weighted standardized difference can be used to compare means, proportions, higher-order moments, and interactions between treated and untreated individuals. In addition, graphical methods can be employed to compare the distribution of continuous covariates between treated and untreated individuals in the weighted population (Austin and Stuart, 2015). Once sufficient covariate balance is achieved, one can estimate the effect of the treatment by direct comparison of outcomes between treated and untreated groups. The weights can also be used in weighted regression models to estimate the effect of the treatment. This method focuses on estimating the average treatment effect in the entire population (ATE); modification of the weights allows to estimate the the average treatment effect in the treated population (ATT) (Stuart, 2010). Most importantly, the variance estimation should take into account the weighted nature of the ”pseudo-population”, for example, by using the sample weights in robust variance estimation (Hernán et al., 2000; Cole and Hernán, 2008). Alternatively, bootstrapping could be used to construct 95% confidence intervals taking into account the estimation of the propensity score(Hernán et al., 2000; Ali et al., 2017b, 2016).

Inverse probability of treatment weights (ITPW) can be also be used to estimate parameters of marginal structural models (MSMs) to deal with time-varying confounding (Hernán et al., 2000), time-modified confounding (Platt et al., 2009), and competing risks (Hernán et al., 2000; Ali et al., 2017b). Hence, the implementation of propensity scores as inverse probability weights is often referred to as MSM using IPTW. All other propensity score approaches can only be extended to time-varying treatment and confounding settings under certain conditions as described below. Comparison of the different propensity score methods is summarized in Table 1.

Table 1 Comparison of the different propensity score methods

Characteristics	Matching[a]	Stratification[b]	Regression[c]	IPTW[d]
Model dependence	Minimum	Minimum	High	Minimum
Application[1]	Easy	Easy	Easy	Complex
Overall transparency	High	High	Low	Medium
Easy to communicate	Yes	Yes	Not always	Not always
Design and analysis	Separated	Separated	Mixed	Separated
Easy to check balance	Yes	Yes	No	Yes
Requires unique assumption[2]	No	No	Yes	No
Excluded subjects from analysis[3]	Yes	No	No	Yes-No
Variance estimation	Not clear	Easy	Easy	Difficult
Easy to interpret[4]	Not always	Yes	No	Often
"Propensity score paradox"	Sensitive	No	No	No
Estimand[5]	Often ATT	ATE, ATT	ATE	ATE, ATT
Time-varying confounding[6]	No	No	No	No/Yes
Multiple treatments	Possible	Complex	complex	Easier
Multi-level treatment applications	Exist	Exist	None	Exist
Treatment effect modification	Easier	Complex	Easier	Complex

[a] Constructs treated and untreated matched groups with similar propensity scores.

[b] Constructs subgroups of treated and untreated subjects, often quintiles or deciles of PS.

[c] PS is used, as a single summary of all covariates included in PS model, in regression model.

[d] PSs are used as weights to create a pseudo-population in which exposure and measured covariates included in the treatment (PS) model are independant.

[1] Estimation of stabilized weights as well as extension to time-varying treatment and confounding setting in MSMs framework can be complex.

[2] Requires correct specification of PS and outcome model, apart from the basic assumptions that there is positivity and no unmeasured confounding.

[3] Weight trimming excludes some individuals in the tails of the propensity score distribution.

[4] In PSM, when treated subject are excluded, interpretation of the treatment effect may change, not just ATT and in Stratification, when there is treatment effect modification by the PS, in regression adjustment using PS, when non-collapsible effect measures such as odds ratios are used.

[5] Modification of the matching or weighting method enable to estimate either ATT or ATE.

[6] When time-varying confounder is affected by previous treatment, all the propensity score based methods fail to correctly control for the confounding bias including standard IPWs; however, MSMs using IPWs can deal with time varying confounding.(Ali et al., 2016)

5 EXTENDED APPLICATIONS

5.1 Time-varying treatments

In clinical practice, it is common for patients to start on a certain medication, stop or switch to another one (for example, due to intolerance or lack of adequate response), in such cases, treatment might be treated as a time-varying exposure. Consider a cohort study to estimate the effect of antiretroviral zidovudine treatment (AZT) in HIV (Human Immunodeficiency Virus) positive individuals, on progression to AIDS (Acquired Immune Deficiency Syndrome), where CD4 count is a confounder. Assuming individuals show up for clinical visits at baseline (t = 0) and then every 6 months (t = 1,2,3, ...), and CD4 counts are recorded at these visits ($CD4_t$, represented as $CD4_0$, $CD4_1$, $CD4_2$, ...). If AZT is a time-varying dichotomous treatment variable indicating whether the individual is on antiretroviral treatment during the interval (t, t+6] months (AZT_t, represented as AZT_0, AZT_1, AZT_2, ...). This means, an individual's treatment plan, at each subsequent visit (t = 1,2,...), is time-varying or dynamic: the clinician in consultation with the individual decide treatment AZT_t based on changing values of individual's clinical and demographic history recorded during the previous and current visits. This includes prior treatment history, current CD4 count, and other confounders which are not included in the DAG and ignored for now for simplicity.

In Figure 2, we considered two time points or visits t = 0 (baseline) and t = 1, hence $CD4_0$ refers to baseline CD4 count and AZT_0 refers to treatment at the first visit. Treatment decision at the first visit AZT_0 is influenced by baseline CD4 count ($CD4_0$), represented in Figure 2A by the arrow from $CD4_0$ to AZT_0. In the second visit (t = 1), treatment decision AZT_1 is based on previous treatment (AZT_0) and CD4 count at the current visit ($CD4_1$), represented in Figure 2A by the arrows from AZT_0 and $CD4_1$ to AZT_1.

In settings such as DAG of Figure 2A, where there is no arrow from AZT_0 to $CD4_1$ implying previous treatment did not affect current CD4 count, all the standard propensity score approaches can deal with the time-varying confounder CD4 count by matching, conditioning, stratification or weighting, for example by using time-varying Cox models. However, this is not biologically plausible since RCTs have proved that antiretroviral treatment indeed affect CD4 count. It is important to mention that there are many practical examples where both treatment and confounders are time-varying or dynamic but previous treatment does not affect time-varying confounder, hence, the DAG in Figure 2A may still be valid.

When a time-varying confounder (such as CD4 count in our example, $CD4_1$) is affected by previous treatment (AZT_0) as in DAG of Figure 2B, the time-varying confounder ($CD4_1$) is also an intermediate for the effect of previous treatment (AZT_0) on the outcome (progression to AIDS), represented by the path AZT_0 $CD4_1$ AIDS). Furthermore, if there exists an unmeasured common cause (U) of both the time-varying confounder ($CD4_1$) and the outcome (progression to AIDS) as in Figure 2C, the time-varying confounder ($CD4_1$) is also a collider on the path AZT_0 $CD4_1$ U AIDS (the arrows collide on $CD4_1$, hence this path is not a causal path because it is blocked at $CD4_1$, in DAG terminologies). That also means there is no association between $CD4_{t-1}$ and U unless we condition, match, or stratify on this collider, $CD4_1$ (Hernán et al., 2000; Robins et al., 2000). Such a time-dependant variable is a confounder, an intermediate, and also a collider at the same time; hence, adjustment requires careful considerations.

Conventional statistical approaches including propensity score methods (matching, stratification, and regression adjustment) that condition or stratify on such a covaiate will result in a biased estimate of the treatment effect (Robins et al., 2000). This happens because conditioning or stratify on an intermediate will adjust-away the indirect effect of the treatment through mediated by the covariate; and conditioning or or stratify on a collider, creates a spurious association between the treatment and the unmeasured common cause that did not exist before conditioning (creating open path $AZT_0 \rightarrow CD4_{t-1} \rightarrow U \rightarrow AIDS$) which is

indicated by using dotted lines in the DAG of Figure 2D, leading to collider-stratification bias (Hernán et al., 2000; Cole et al., 2009).

In such settings, MSM using inverse probability weighting is the method of choice; unlike conditioning or stratification, weighting creates a ”psudo-population” in which the association between the time-varying confounder and treatment is removed (Robins et al., 2000). Additional methods are available to deal with time-varying treatment and confounding including other classes of marginal structural models (g-formula and g-estimation of structural nested models)(Robins et al., 2000; Hernán et al., 2000).

It is straightforward to hypothesize that such a time-varying confounding can also be time-modified, which means not only the confounder (CD4 count) change over time but also its impact on the outcomes (progression to AIDS) varies during these times. The effects of the confounder change over time means that the strength of association between $CD4_0$ and AIDS ($CD4_0 \rightarrow$ AIDS) is different from that of $CD4_1$ and AIDS ($CD4_0 \rightarrow$ AIDS) Platt et al. (2009). However, time-modified confounding might still exist in longitudinal treatment settings where the confounder is time-invariant or fixed. Standard methods are sufficient to deal with time-modified confounding unless the the confounders are both time-varying and affected by previous treatment, which requires the use of marginal structural models, such as the use of inverse probability weighting.

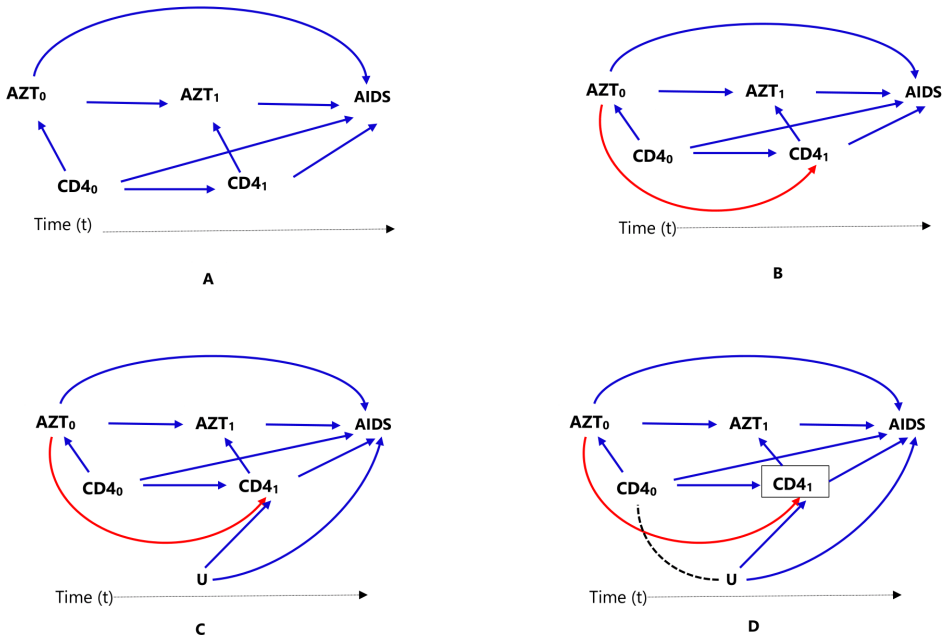

Figure 2. Causal diagrams representing time-varying treatment (AZT), outcome (progression to AIDS, AIDS), time-varying confounding (CD4 count). Time-varying confounding is not affected by prior treatment (A), time-varying confounding is affected by prior treatment (B), time-varying confounding affected by unmeasured factor U which is also associated with the outcome (C), and conditioning or stratifying on time-varying confounder, indicated by box around $CD4_1$ creates association between time varying confounder $CD4_0$ and unmeasured factor U (D).

5.2 Multiple treatments

Propensity score is often used to estimate the effect of a binary treatment (whether treatment is received: Yes = 1 or No = 0) in observational data. However, with more than two levels of treatment which is common in pharmacoepidemiology such as comparison of three or more statins (e.g., simvastatin, atorvastatin, fluvastatin, lovastatin, pravastatin, and rosuvastatin) or of multiple doses of a certain medication (e.g., low- medium-high doses), estimation of treatment effects requires additional assumptions and modelling techniques (Imbens, 2000; McCaffrey et al., 2004). These include the use of multinomial logistic and multinomial probit models for nominal treatments and ordinal logistic regression or the proportional odds model for ordinal treatments (Imbens, 2000). Alternatively, generalized boosted model, a machine learning approach involving an iterative process with multiple regression trees to capture complex and non-linear relationships between treatment assignment and pre-treatment covariates without over-fitting the data, can be used to fit inverse probability weighting for multiple treatments (McCaffrey et al., 2004). However, applications in pharmacoepidemiology using observational data are infrequent partly due to methodological complexities in fitting the models and understanding their assumptions as well as limited availability of guidance documents on the methods.

5.3 Multi-level treatments

Propensity score methods have been extensively studied and widely applied in a single-level treatment (no clustering among participants); however, most health care data have a multi-level structure such that subjects are grouped into clusters such as geographical area, treatment center (hospital or physician), or insurance plans (Goldstein et al., 2002). The unknown mechanisms that assigns subjects to clusters may be associated with individual-level measured confounders (such as race, age, and clinical characteristics) and unmeasured confounders(such as unmeasured severity of disease, aggressiveness in seeking treatment). These measured and unmeasured confounders might also create a cluster-level variation in treatment and/or outcomes. If this variation is correlated with group assignment at the group or cluster level, it might lead to confounding (Greenland, 2000). hence, the use of standard regression or propensity score methods ignoring the cluster structure should should be avoided. This is because ignoring the cluster structure often leads to invalid inferences: not only the standard errors are inaccurate but also the cluster-level effects could be confounded with individual-level effects.

Propensity score matching and weighting are often used in such settings (Arpino and Mealli, 2011; Li et al., 2013). One might consider the use of within-cluster PSM (of treated and untreated subjects) which automatically achieve perfect balance on all the measured and measured cluster characteristics. However, it is very unlikely, particularly in small clusters, to find sufficient number of untreated matches to treated subjects in the same cluster. Alternatively, PSM could be performed across clusters taking into account the cluster structure in the propensity score estimation model. Preferably, cluster structure should be taken in to account in estimation of both the propensity score and the treatment effect (Li et al., 2013).

Multilevel regression models that include fixed effects and/or random effects have been developed (Greenland, 2000; Goldstein et al., 2002), and extended to propensity scores approaches (Arpino and Mealli, 2011). Empirical applications of such methods in medication and device effectiveness and safety are rare. However, simulations studies have shown that multi-level propensity score matching (Arpino and Mealli, 2011) and weighting approaches (Li et al., 2013), without imposing a within-cluster matching or weighting requirement, reduce bias due to unmeasured cluster-level confounders .

5.4 Propensity score with missing data

Missing data is a common problem in the estimation of treatment effects using routinely collected data. The impact of such missing data on the results of the treatment effect estimation depends on the mechanism which caused the data to be missing and the way missing data is handled. Missing data can be categorized in to three distinct classes based on the relationship between the missing data mechanism and the missing and observed values: i) Missing Completely at Random (MCAR) when the missing data mechanism is unrelated to the values of any variable, whether missing or observed. Hence, the observed values are representative of the entire sample without missing values. ii) Missing at Random (MAR) when the missing data mechanism is unrelated to the missing values but may be related to the observed values of other variables.iii). Missing not at Random (MNAR) when the missing data mechanism is related not only to the observed values of other variables but also to the missing values (Rubin, 1996). For each of the missing data patterns, different statistical techniques are used to handle its impact on the quality the inference. It is important to emphasize that MCAR, MAR, and MNAR could exist for different variables in a specific data.

Complete case analysis, including only those individuals who have no missing data in any of the variables that are required for the analysis, does well when data is missing completely at random (MCAR). However, it may result in biased estimate of the treatment effect if missing is at random (MAR) (Rubin, 1996; Sterne et al., 2009). In MAR, as stated before, any systematic difference between the missing values of a variable and the observed values of the variable can be explained by differences in observed data (Sterne et al., 2009). Furthermore, missing data in several variables often leads to exclusion of a substantial proportion of the original sample, which in turn causes a substantial loss of precision (i.e., power) and hence results in wider confidence intervals (Cummings, 2013). Other approaches to deal with missing data including replacing missing values with values imputed from the observed data (for example, the mean of the observed values), using a missing category indicator, and replacing missing values with the last measured value particularly in longitudinal studies (”last value carried forward”) are generally statistically invalid and they could lead to serious bias (Rubin, 1996; Sterne et al., 2009). In addition, single imputation of missing values usually results in too small standard errors, because it fails to account for the uncertainity about the missing values (Sterne et al., 2009).

A relatively flexible approach to allow for the uncertainty in the missing data is multiple imputation. Multiple imputation involves creating multiple different copies of the dataset with the missing values replaced by imputed values (Step 1); estimating treatment effects in each copy of the data (Step 2); averaging the estimated treatment effects to give overall estimated measure of association and calculating standard errors using Rubin’s rules (Step 3) (Rubin, 1996, 2004a). Applications of propensity score methods in data with missing values involve a similar approach: 1) Creation of multiple copies of imputed data. 2) Estimation of propensity scores and treatment effects in each of the imputed copies of the dataset (Qu and Lipkovich, 2009; Leyrat et al., 2019). 3) Pooling of treatment effects by averaging across the multiple data sets and estimation of standard errors using Ruben’s rule (Crowe et al., 2010; Leyrat et al., 2019) (Figure 3A). An alternative approach is pooling the propensity scores from the multiple copies of data, in step 2, and conducting the analysis in the pooled data (Figure 3B); however, this method has been proved sub-optimal in terms of bias reduction (Leyrat et al., 2019).

6 ADVANTAGES AND LIMITATIONS OF PROPENSITY SCORE METHODS

Previous literature reviews of observational studies have found that results from both traditional regression and propensity scores analyses are similar (Shah et al., 2005; Stürmer et al., 2006a). These findings may

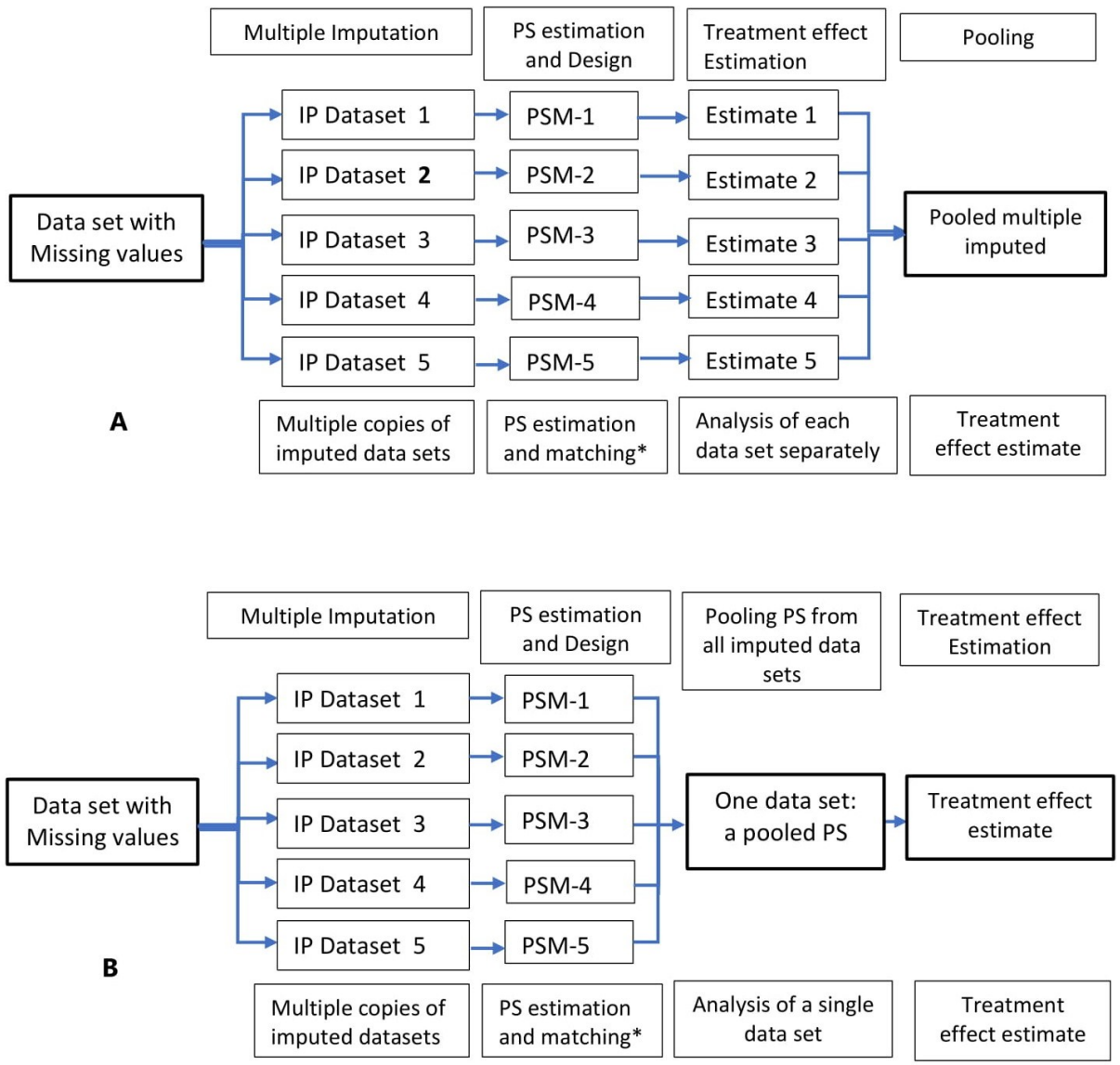

Figure 3. Multiple imputation in propensity score methods, multiple copies of imputed data are created and propensity score is estimated using these data sets.Treatment effects are estimated in several data sets and pooled (A) propensity scores from multiple data sets are pooled and treatment effect estimated in a single data set(B).* Other PS methods, stratification, IPTW, and covariate adjustment using PS could also be used instead of matching.

be in part due to sub-optimal implementations of propensity score methods (Shah et al., 2005; Austin, 2008a; Ali et al., 2015); however, similarity of findings have been used to question the need for propensity score methods if they do not provide better ways to improve confounding control. Despite these findings, propensity score methods will remain advantageous for several reasons compared to covariate-adjustment techniques, which correct for covariate imbalances between treatment groups by controlling for them in regression models for the outcomes.

Transparency

Propensity score methods primarily aim at balancing treatment groups with respect to covariate distributions; when such balance is achieved, it is relatively easy to detect and communicate (Ali et al., 2015) by using simple graphical tools or quantitative statistics. In addition, propensity score methods, unlike regression adjustment, can give investigators an insight into the quality of the data at hand. Inadequate overlap in propensity score distributions (i.e., poor "common support") between treatment groups should be considered as a warning that the particular dataset cannot address the causal question without relying on untrustworthy "model-dependent" extrapolations (Rubin, 2004b, 2007; Dehejia and Wahba, 2002). In some case, the researcher might decide to focus on subjects only in the overlapping regions using propensity score matching or trimming; as consequence, the conclusions of the findings should be restricted to subjects that are sufficiently represented in the overlapping regions of the propensity score distributions (Ali et al., 2016). Conventional regression methods do not provide the researcher with these possibilities. Furthermore, covariate balance in regression methods is a "black-box" and, irrespective of inadequate overlap (i.e., when the treated and untreated groups are disparate on pre-treatment covariates), conventional models use extrapolations to estimate treatment effects that may not be generalizable to the entire population in the data.

Design tools

Similar to RCTs, propensity score methods can be considered as design tools for pre-processing of the data (matching, stratification, and weighting) without using any outcome information at this stage. As a result, formal causal inference models (also called the potential outcomes framework) (Rubin, 2005) can be applied to clearly specify the causal question without conflating with the modeling approach (Vandenbroucke et al., 2016), hence it allows for a simple and transparent analysis. In addition, this approach minimizes bias from potential misspecification of the outcome model (Rubin, 2004b). Furthermore, matched, stratified, and weighted analyses do not make strong assumptions of linearity in the relationship of propensity score with the outcome. If a non-parametric pre-processing of the data using propensity score methods does not reduce model-dependence, it is likely that the data contain little information to reliably support the causal inference by any other method. Obviously, this knowledge in itself would still be useful information and the conclusion may be correct (Rubin, 2004b; Ho et al., 2007; Rubin, 2007).

Dimension reduction

Propensity score typically summarizes large number of measured pre-treatment covariates to a single score, hence it is called a "summary score". This is particularly useful in high dimensional data with substantially large number of pre-treatment covariates compared to the number of outcome events including rare events, typical of most medication safety studies in pharmacoepidemiology (Glynn et al., 2006). In this setting, maximum likelihood estimations used in conventional regression techniques such as logistic and Cox regression requires many outcomes per included parameter in a model, the rule of thumb is that $\geq$ 10 outcome events are required for every covariate included in a regression model (Peduzzi et al., 1995, 1996). On the other hand, Cepeda et al. (Cepeda et al., 2003) suggested using propensity score when there are fewer than eight outcomes per included covariate to effectively improve estimation.

Doubly robust estimations

Generally, doubly robust estimation methods apply different procedures or models simultaneously and produce a consistent estimate of the parameter if either of the two models, not necessarily both, has

been correctly specified (Imai and Ratkovic, 2014). Several applications of propensity scores have been described as doubly robust in terms of estimating the effect of a certain treatment, including:

1) The combined use of propensity score methods (matching, regression, or weighting) with regression adjustments. These approaches use non-parametric pre-processing of the data to minimizing imbalances in measured covariates and, if there are still residual differences, the covariates can be adjusted in the outcome model (Rubin and Thomas, 2000; Nguyen et al., 2017).

2) The combined use of propensity and prognostic score methods (Leacy and Stuart, 2014; Ali et al., 2018b), a prognostic score is as any function of a set of covariates that when conditioned on induces independence between the potential outcome under the control (no treatment) condition and the unreduced covariates (Hansen, 2008). Hence, differences in outcomes between treated and untreated subjects can be attributed to the effect of the treatment under study. The two approaches could be combined in several ways such as full matching on a Mahalanobis distance combining the estimated propensity and prognostic scores; full matching on the estimated prognostic score within propensity score calipers; and sub-classification on an estimated propensity and prognostic score grid with 5 5 sub-classes, among others (Leacy and Stuart, 2014; Ali et al., 2018b). Methods combining propensity and prognostic scores were no less robust to model misspecification than single-score methods even when both prognostic and propensity score models were incorrectly specified in simulation and empirical studies (Leacy and Stuart, 2014).

3) The use of covariate balancing propensity score (CBPS) introduced by Imai et al. (Imai and Ratkovic, 2014) which involves estimation of the propensity score such that the resulting covariate balance is optimized. This approach utilizes the dual characteristics of the propensity score as a covariate balancing score and the conditional probability of treatment assignment. Specifically, the covariate balancing property (i.e., mean independence between the treatment status and measured covariates after inverse propensity score weighting) is used as conditions to imply estimation of the propensity score while also incorporating the standard estimation procedure. Unlike other covariate balancing methods, a single model determines the treatment assignment mechanism and the covariate balancing weights. Once covariate balancing propensity score is estimated, various propensity score methods such as matching and weighting can be implemented without modification (Imai and Ratkovic, 2014).

4) Calculation of doubly robust (DR) estimator using the propensity score, predicted and observed outcome ($\hat{Y}$ and Y, respectively). It involves specifying regression models for the treatment (Z) and the outcome (Y) as a function of covariates (X) and combining these subject-specific values to calculate DR estimate for each individual. First, treatment is modelled as a function of covariates to estimate propensity scores for each individual using the observed data. Second, the relationships between measured confounders and the outcome is modelled within treated and untreated groups separately. The resulting parameter estimates are then used to calculate predicted outcomes ($\hat{Y}_1$, $\hat{Y}_0$) for each individual in the population that is treated (setting Z = 1) and not treated (setting Z = 0) given covariate values. Third, the doubly robust (DR) estimates of the outcome is calculated for each individual both in the presence and absence of treatment (DR_1 and DR_0, respectively) using the subject-specific predicated ($\hat{Y}$) and observed (Y) outcomes weighted by the propensity score. Finally, the means of DR_1 and DR_0 are calculated across the entire study population and these means will be used to calculate the effect of the treatment (Funk et al., 2011).

Unmeasured confounding

Propensity score methods, alike other conventional regression methods, can control for only measured confounding factors and not unmeasured ones (Rosenbaum and Rubin, 1983). As a result, propensity score

663 analysis can only be as good as the quality and the completeness of potential confounding variables that
664 are at the disposal of the researcher. Only a rich set of covariates may convince a critical reader that no 665
unmeasured confounding variables were missed. Therefore, it is important to provide a detailed account of 666
the variables collected and included in the propensity score model (Ali et al., 2015).

667 Modifications of the standard propensity score applications have been suggested to further reduce the risk
668 of unmeasured confounding including the use of high dimensional propensity score and propensity score
669 calibration. High dimensional propensity score refers to the use of large number (in the range of several 670
hundreds) of covariates to improve control of confounding; the underlying assumption is that the variables 671
may collectively be proxies for unobserved confounding factors (Schneeweiss et al., 2009; Rassen et al., 672
2011). Propensity score calibration refers to the use of a ”gold-standard” propensity score estimated in 673 a
separate validation study, with more detailed covariate information unmeasured in the main study, to 674
correct the main-study effect of the drug on the outcome (Stürmer et al., 2005, 2007).

675 Furthermore, sensitivity analyses (Rosenbaum and Rubin, 1983; Rosenbaum, 2005) are useful to assess 676
the plausibility of the assumptions underlying the propensity score methods and how violations of them 677
might affect the conclusions drawn (Stuart, 2010). Methods to deal with unmeasured confounding are 678
summerized in Figure 1.

679 Effect modification

680 In estimating treatment effects, there is often an interest to explore if the effect of treatment varies 681
among different subgroups (for example, men versus women) of the population under study, often called 682
”treatment effect modification”. There are many ways to utilize propensity score methods to adjust for 683
confounding in a subgroup analysis; however, common implementation of propensity score matching in 684 the
medical literature are sub-optimal (Wang et al., 2017; Ali et al., 2018a). The use of propensity score 685 matched
(PSM) cohort for subgroup analysis breaks the matched sets and might result in imbalance of 686 covariates
(Ali et al., 2018a). Depending on the frequency of treatment or outcome, small changes in the 687 matched
cohort can result in large fluctuations for measures of association (Rassen et al., 2012).

688 To account for covariate imbalances, subgroup analyses of propensity score matched cohorts involve: 689
i) adjusting for covariates in the outcome model or ii) re-matching within the subgroups either using the 690
propensity score estimated in the full cohort or fitting new propensity score within subgroups (Figure 4) 691
(Rassen et al., 2012; Wang et al., 2017). The choice of a specific method should take in to account several 692
factors: prevalence of treatment and outcome; strength of association between pre-treatment covariates and 693
treatment; the true effect within subgroups, and the degree of confounding within subgroups (Wang et al., 694
2018).

7 REPORTING

695 The credibility of any research depends on a critical assessment by others of the strengths and weaknesses
696 in study design, conduct, and analysis. Hence, transparent and adequate reporting of critical aspects of 697
propensity score-based analysis (Ali et al., 2015), like other observational studies, helps readers follow 698
”what was planned, what was done, what was found, and what conclusions were drawn” (Von Elm et al., 699
2007). It also makes it easier for other researchers to replicate the study findings using other data sources 700
and to judge whether and how results can be included in systematic reviews (Von Elm et al., 2007). Despite 701
substantial developments and common applications of propensity score methods, reporting on aspects of 702
the propensity score analysis is generally poor and inconsistent in the medical literature (Austin, 2008a; Ali

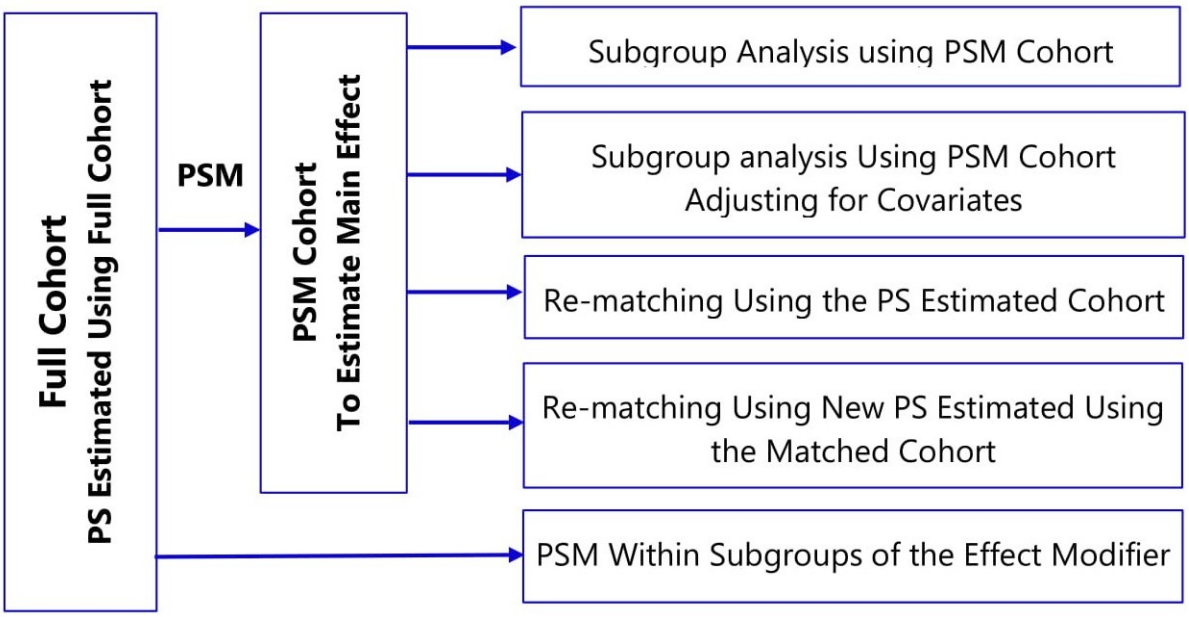

Figure 4. Methods to assess treatment effect modification in propensity score matching.

et al., 2015, 2016; Wang et al., 2017). This could in part be due to a lack of standards for conduct as well as reporting of propensity score methods in guidelines. Therefore, critical items relevant to propensity score analysis should be incorporated in guidelines on the conduct and reporting of observational studies, such as the STROBE statement (Von Elm et al., 2007) and the ENCePP guide on methodological standards in pharmacoepidemiology (Blake et al., 2012) to improve the quality of conduct and reporting of propensity score based studies (Ali et al., 2015, 2016). Table 2 summerizes important consideration when planning, conducting, and reporting propensity score analysis and list of items that should be reported is summarized by Ali et al.(Ali et al., 2016).

Table 2 Summary of considerations when planning, conducting, and reporting propensity score analysis.

Characteristics	What to consider	Methods available to deal with	What should or should not be done
Missing data	Missing data mechanism	Multiple imputation if missing at random (MAR)	Complete case analysis, mean imputation, or missing indicator category (if missing is not completely at random, MCAR)
Variable selection	Potential confounders, intermediates, colliders	Clinical knowledge/expert opinion.	Avoid adjusting for intermediates, colliders, and strong instrumental variables (only when sure and suspect unmeasured confounding).
		Association between variables with outcome (and treatment).	Avoid the use of p-values, or step-wise variable selection methods.
		Balance diagnostics.	
Propensity score estimation	Variables included, interactions and higher order terms.	Logistic regression, Recursive partitioning, Neural network, Classification and regression trees, Random forest, and Boosting regression.	Report on the method used for estimation and variables included in the propensity score method.
Propensity score methods	The research question, the treatment effect estimand, and the extent of overlap.	Density plots of propensity scores.	Report the density plots or histograms in the propensity score distribution (preferably overlapping coefficients of the density plots).

Propensity score matching	Matching algorithm, matching with or without replacement, and matching ratio	Exact (coarsened) matching, nearest neighbour matching (with or without calliper), stratified matching, and full matching. Matching ratio can be: 1-to-1 matching, 1-to- many matching, variable ratio matching, and full matching.	Report on the number of starting population, number matched, and number excluded (with their baseline characteristics).
Propensity score stratification	Number of strata	Deciles and quintiles of propensity scores.	Report on the number of strata used and the covariate balance between treatment groups in each strata.
Regression adjustment using propensity score	Linear relationship between the outcome and the propensity core.		Report on whether linear relationship between the outcome and propensity core is checked and is fulfilled.
Inverse probability of treatment weighting	Whether there is sufficient overlap (positivity).	Weighted regression.Robust variance estimation or Bootstrapping for constructing confidence intervals.	Report on how weights are calculated, if weights are stabilized, the mean weights in both treatment groups, if trimming has been done.
Time-varying exposure	Whether there is time-varying confounding, and if any, whether it is affected by previous treatment.	Marginal Structural models using IPTW, G-formula and G-estimation of structural nested models.	If previous treatment affect time-varying confounding avoid matching, stratification and regression adjustment; use MSM using IPTW.
Treatment effect modification	Identify potential effect modifier.	Matching on PS within strata of effect modifier, among others.	Avoid the use of stratified analysis using the PSM data without adjustment for covariates.

Multilevel treatment	Whether multi-level structure exist in the data, the number of clusters/levels.	Multi-level propensity score methods.	Avoid use of single-level propensity score applications. Include multi-level structure at least in propensity score estimation or outcome analysis, preferably in both.
Multiple treatments	Number of treatment groups, whether there is order in the treatment categories (such as dosage).	Multiple matching and weighting: multinomial logistic regression, ordinal logistic regression, or generalized boosted model.	
Residual Confounding	Whether there is imbalance in covariates.	Doubly robust methods, propensity score calibration, high dimensional propensity score method.	Report on which method was used and why?
Unmeasured confounding	Whether there is potential unmeasured confounding, or whether the data contain proxies for unmeasured confounding.	Alternative methods such as instrumental methods, propensity score calibration, or consider sensitivity analysis.	Report on the method used and the sensitivity analysis conducted.

8 CONCLUSION

Propensity score methods will remain important design and analytic tools to estimate effects of treatment from observational data. Preferably, they should be utilized in the design stage as tools for pre-processing of the data and, when appropriate, in combination with model-based adjustment methods. They should neither be regarded as a "panacea" for the deficiencies of observational studies such as residual or unmeasured confounding nor as replacement for conventional regression adjustments (Ali et al., 2016). The ability of propensity score methods to overcome confounding is entirely dependent on the extent to which measured variables capture potential confounding. Taking full advantage of these methods requires explicit definition of the research question and appropriate choice of the propensity score method, transparent and detailed description of all subsequent statistical analyses to be conducted, and adequate reporting of important aspects of the propensity score analysis (Ali et al., 2016).

CONFLICT OF INTEREST STATEMENT

LS holds research grants from GSK, Wellcome, MRC, NIHR, BHF, Diabetes UK and Is a Trustee of the British Heart Foundation. DPA holds research grants from NIHR, AMGEN, UCB. Others have no conflict of interest.

AUTHOR CONTRIBUTIONS

MSA, DPA, RF, MLB and LS contributed conception and design of the study; MSA wrote the first draft of the manuscript; DR and NB wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

The is no specific funding for this work.

ACKNOWLEDGMENTS

REFERENCES

Adelson, J. L., McCoach, D., Rogers, H., Adelson, J. A., and Sauer, T. M. (2017). Developing and applying the propensity score to make causal inferences: variable selection and stratification. *Frontiers in psychology* 8, 1413

Ali, M. S., Collins, G., and Prieto-Alhambra, D. (2017a). The "propensity score paradox": A threat to pharmaco-epidemiological studies?

Ali, M. S., Douglas, I. J., Williamson, E., Prieto-Alhambra, D., and Smeeth, L. (2018a). Evaluation of treatment effect modification in propensity score matching: An empirical example. In *Pharmacoepidemiology and drug safety* (WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA), vol. 27, 25–25

Ali, M. S., Douglas, I. J., Williamson, E., Prieto-Alhambra, D., and Smeeth, L. (2018b). A joint application of disease risk score and propensity score to control for confounding: A clinical example. In *Pharmacoepidemiology and drug safety* (WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA), vol. 27, 27–27

741 Ali, M. S., Groenwold, R. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C., et al. (2015). 742 Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a 743 systematic review. *Journal of clinical epidemiology* 68, 122–131

744 Ali, M. S., Groenwold, R. H., and Klungel, O. H. (2016). Best (but oft-forgotten) practices: propensity 745 score methods in clinical nutrition research–3. *The American journal of clinical nutrition* 104, 247–258 746 Ali, M. S., Groenwold, R. H., Pestman, W. R., Belitser, S. V., Roes, K. C., Hoes, A. W., 747 et al. (2014). Propensity score balance measures in pharmacoepidemiology: a simulation study. 748 *Pharmacoepidemiology and drug safety* 23, 802–811

749 Ali, M. S., Khalid, S., Collins, G., and Prieto-Alhambra, D. (2017b). The comparative performance of 750 logistic regression and random forest in propensity score methods: A simulation study

751 Ali, M. S., Khalid, S., Groenwold, R., Collins, G. S., Klungel, O., and Prieto-Alhambra, D. (2017c). 752 Instrumental variables to test for unmeasured confounding: a precautionary note

753 Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental 754 variables. *Journal of the American statistical Association* 91, 444–455

755 Arpino, B. and Mealli, F. (2011). The specification of the propensity score in multilevel observational 756 studies. *Computational Statistics & Data Analysis* 55,1770–1780

757 Austin, P. C. (2008a). A critical appraisal of propensity-score matching in the medical literature between 758 1996 and 2003. *Statistics in medicine* 27,2037–2049

759 Austin, P. C. (2008b). The performance of different propensity-score methods for estimating relative risks. 760 *Journal of clinical epidemiology* 61, 537–545

761 Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between 762 treatment groups in propensity-score matched samples. *Statistics in medicine* 28, 3083–3107

763 Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding 764 in observational studies. *Multivariate behavioral research* 46,399–424

765 Austin, P. C. (2019). Assessing covariate balance when using the generalized propensity score with 766 quantitative or continuous exposures. *Statistical methods in medical research* 28, 1365–1377

767 Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of 768 treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational 769 studies. *Statistics in medicine* 34, 3661–3679

770 Bärnighausen, T., Tugwell, P., Røttingen, J.-A., Shemilt, I., Rockers, P., Geldsetzer, P., et al. (2017). 771 Quasi-experimental study designs series—paper 4: uses and value. *Journal of clinical epidemiology* 89, 772 21–29

773 Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H., De Boer, A., and Klungel, O. H. (2011). 774 Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and drug* 775 *safety* 20, 1115–1129

776 Blake, K. V., deVries, C. S., Arlett, P., Kurz, X., Fitt, H., and of Centres for 777 Pharmacoepidemiology Pharmacovigilance, E. N. (2012). Increasing scientific standards, independence 778 and transparency in post-authorisation studies: the role of the european network of centres for 779 pharmacoepidemiology and pharmacovigilance. *Pharmacoepidemiology and drug safety* 21, 690–696 780 Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable 781 selection for propensity score models. *American journal of epidemiology* 163, 1149–1156

782 Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. (2003). Comparison of logistic regression versus 783 propensity score when the number of events is low and there are multiple confounders. *American journal* 784 *of epidemiology* 158, 280–287

785 Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural
786 models. *American journal of epidemiology* 168, 656–664
787 Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., et al. (2009).
788 Illustrating bias due to conditioning on a collider. *International journal of epidemiology* 39, 417–420
789 Concato, J., Shah, N., and Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and
790 the hierarchy of research designs. *New England Journal of Medicine* 342, 1887–1892
791 Crowe, B. J., Lipkovich, I. A., and Wang, O. (2010). Comparison of several imputation methods for
792 missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical statistics* 9,
793 269–279
794 Cummings, P. (2013). Missing data and multiple imputation. *JAMA pediatrics* 167, 656–661
795 Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal
796 studies. *Review of Economics and statistics* 84, 151–161
797 D'Agostino Jr, R. B. (2007). Propensity scores in cardiovascular research. *Circulation* 115, 2340–2343
798 Eichler, H.-G., Abadie, E., Breckenridge, A., Flamion, B., Gustafsson, L. L., Leufkens, H., et al. (2011).
799 Bridging the efficacy–effectiveness gap: a regulator's perspective on addressing variability of drug
800 response. *Nature reviews Drug discovery* 10, 495
801 Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly
802 robust estimation of causal effects. *American journal of epidemiology* 173, 761–767
803 Glynn, R. J., Schneeweiss, S., and Stürmer, T. (2006). Indications for propensity scores and review of their
804 use in pharmacoepidemiology. *Basic & clinical pharmacology & toxicology* 98, 253–259
805 Goldstein, H., Browne, W., and Rasbash, J. (2002). Multilevel modelling of medical data. *Statistics in*
806 *medicine* 21, 3291–3315
807 Greenland, S. (2000). Principles of multilevel modelling. *International journal of epidemiology* 29,
808 158–167
809 Greenland, S. and Morgenstern, H. (2001). Confounding in health research. *Annual review of public health*
810 22, 189–212
811 Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and collapsibility in causal inference.
812 *Statistical science* , 29–46
813 Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the*
814 *American Statistical Association* 99, 609–618
815 Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika* 95, 481–488
816 Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal
817 effect of zidovudine on the survival of hiv-positive men. *Epidemiology* , 561–570
818 Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An
819 application to data on right heart catheterization. *Health Services and Outcomes research methodology*
820 2, 259–278
821 Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for
822 reducing model dependence in parametric causal inference. *Political analysis* 15, 199–236
823 Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical*
824 *Society: Series B (Statistical Methodology)* 76, 243–263
825 Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the
826 propensity score. *Journal of the American Statistical Association* 99, 854–866
827 Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*
828 87, 706–710

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics* 86, 4–29

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press)

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: propensity scores. *American journal of epidemiology* 150, 327–333

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., and Feldman, B. M. (2018). Propensity score methods for bias reduction in observational studies of treatment effect. *Rheumatic Disease Clinics* 44, 203–213

King, G. and Nielsen, R. (2016). Why propensity scores should not be used for matching. *Copy at http://j.mp/1sexgVw Download Citation BibTex Tagged XML Download Paper* 378

Leacy, F. P. and Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in medicine* 33, 3488–3508

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine* 29, 337–346

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one* 6, e18174

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., et al. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical methods in medical research* 28, 3–19

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

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in medicine* 32, 3373–3387

Linden, A. and Samuels, S. J. (2013). Using balance statistics to determine the optimal number of controls in matching studies. *Journal of evaluation in clinical practice* 19, 968–975

Lunt, M. (2013). Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. *American journal of epidemiology* 179, 226–235

Lunt, M., Solomon, D., Rothman, K., Glynn, R., Hyrich, K., Symmons, D. P., et al. (2009). Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *American journal of epidemiology* 169, 909–917

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

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., et al. (2011a). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology* 174, 1213–1222

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., et al. (2011b). Myers et al. respond to "understanding bias amplification". *American journal of epidemiology* 174, 1228–1229

Nguyen, T.-L., Collins, G. S., Spence, J., Daurès, J.-P., Devereaux, P., Landais, P., et al. (2017). Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC medical research methodology* 17, 78

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., et al. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiology and drug safety* 20, 551–559

874 Pearl, J. (2011). Invited commentary: understanding bias amplification. *American journal of epidemiology*
875 174, 1223–1227
876 Pearl, J. (2012). On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint*
877 *arXiv:1203.3503*
878 Peduzzi, P., Concato, J., Feinstein, A. R., and Holford, T. R. (1995). Importance of events per independent
879 variable in proportional hazards regression analysis ii. accuracy and precision of regression estimates.
880 *Journal of clinical epidemiology* 48, 1503–1510
881 Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of
882 the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology* 49,
883 1373–1379
884 Platt, R. W., Schisterman, E. F., and Cole, S. R. (2009). Time-modified confounding. *American journal of*
885 *epidemiology* 170, 687–694
886 Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple
887 imputation missingness pattern (mimp) approach. *Statistics in Medicine* 28, 1402–1414
888 Rassen, J. A., Glynn, R. J., Brookhart, M. A., and Schneeweiss, S. (2011). Covariate selection in high-
889 dimensional propensity score analyses of treatment effects in small samples. *American journal of*
890 *epidemiology* 173, 1404–1413
891 Rassen, J. A., Glynn, R. J., Rothman, K. J., Setoguchi, S., and Schneeweiss, S. (2012). Applying propensity
892 scores estimated in a full cohort to adjust for confounding in subgroup analyses. *Pharmacoepidemiology*
893 *and drug safety* 21, 697–709
894 [Dataset] Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal
895 inference in epidemiology
896 Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. *Encyclopedia of statistics in*
897 *behavioral science* 4, 1809–1814
898 Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies
899 for causal effects. *Biometrika* 70, 41–55
900 Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification
901 on the propensity score. *Journal of the American statistical Association* 79, 516–524
902 Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched
903 sampling methods that incorporate the propensity score. *The American Statistician* 39, 33–38
904 Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*
905 91, 473–489
906 Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of*
907 *internal medicine* 127, 757–763
908 Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the
909 tobacco litigation. *Health Services and Outcomes Research Methodology* 2, 169–188
910 Rubin, D. B. (2004a). *Multiple imputation for nonresponse in surveys*, vol. 81 (John Wiley & Sons)
911 Rubin, D. B. (2004b). On principles for modeling propensity scores in medical research.
912 *Pharmacoepidemiology and drug safety* 13, 855–857
913 Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of*
914 *the American Statistical Association* 100, 322–331
915 Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels
916 with the design of randomized trials. *Statistics in medicine* 26, 20–36
917 Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to
918 practice. *Biometrics* , 249–264

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

Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods* 13, 279

Schneeweiss, S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and drug safety* 15, 291–303

Schneeweiss, S., Gagne, J., Glynn, R., Ruhl, M., and Rassen, J. (2011). Assessing the comparative effectiveness of newly marketed medications: methodological challenges and implications for drug development. *Clinical Pharmacology & Therapeutics* 90, 777–790

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)* 20, 512

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

Shah, B. R., Laupacis, A., Hux, J. E., and Austin, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of clinical epidemiology* 58, 550–559

Sibbald, B. and Roland, M. (1998). Understanding controlled trials. why are randomised controlled trials important? *BMJ: British Medical Journal* 316, 201

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* 338, b2393

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of 'a critical appraisal of propensity score matching in the medical literature between 1996 and 2003'by peter austin, statistics in medicine. *Statistics in medicine* 27, 2062–2065

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., and Schneeweiss, S. (2006a). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of clinical epidemiology* 59, 437–e1

Stürmer, T., Rothman, K. J., and Glynn, R. J. (2006b). Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiology and drug safety* 15, 698–709

Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American journal of epidemiology* 162, 279–289

Stürmer, T., Schneeweiss, S., Rothman, K. J., Avorn, J., and Glynn, R. J. (2007). Performance of propensity score calibration—a simulation study. *American journal of epidemiology* 165, 1110–1118

Tarricone, R., Boscolo, P. R., and Armeni, P. (2016). What type of clinical evidence is needed to assess medical devices? *European Respiratory Review* 25, 259–265

Uddin, M. J., Groenwold, R. H., Ali, M. S., de Boer, A., Roes, K. C., Chowdhury, M. A., et al. (2016). Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *International journal of clinical pharmacy* 38, 714–723

964 Vandenbroucke, J. P., Broadbent, A., and Pearce, N. (2016). Causality and causal inference in epidemiology:
965 the need for a pluralistic approach. *International journal of epidemiology* 45, 1776–1786
966 Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., et al. (2007).
967 The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines 968
for reporting observational studies. *PLoS medicine* 4, e296
969 Wang, S. V., He, M., Jin, Y., Wyss, R., Shin, H., Ma, Y., et al. (2017). A review of the performance of 970
different methods for propensity score matched subgroup analyses and a summary of their application in 971
peer-reviewed research studies. *Pharmacoepidemiology and drug safety* 26, 1507–1512
972 Wang, S. V., Jin, Y., Fireman, B., Gruber, S., He, M., Wyss, R., et al. (2018). Relative performance of
973 propensity score matching strategies for subgroup analyses. *American journal of epidemiology*
974 Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., and Mor, V. (2005). Weaknesses of 975
goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. 976
Pharmacoepidemiology and drug safety 14, 227–238
977 Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A., and Stürmer, T. (2011). The role of the c-statistic
978 in variable selection for propensity score models. *Pharmacoepidemiology and drug safety* 20, 317–320 979
Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support 980
vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal* 981 *of*
clinical epidemiology 63, 826–833
982 Yuan, H., Ali, M. S., Brouwer, E. S., Girman, C. J., Guo, J. J., Lund, J. L., et al. (2018). Real-world evidence:
983 What it is and what it can tell us according to the international society for pharmacoepidemiology
984 (ispe) comparative effectiveness research (cer) special interest group (sig). *Clinical Pharmacology &*
985 *Therapeutics* 104, 239–241
986 Zhang, X., Faries, D. E., Li, H., Stamey, J. D., and Imbens, G. W. (2018). Addressing unmeasured
987 confounding in comparative observational research. *Pharmacoepidemiology and drug safety* 27, 373–
988 382