Best (but oft-forgotten) practices: propensity score methods in clinical nutrition research[1–3]

*M Sanni Ali,[4–6] Rolf HH Groenwold,[5,6] and Olaf H Klungel[5,6]**

[4]Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom; [5]Utrecht Institute for Pharmaceutical Sciences, Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht University Utrecht, Netherlands; and [6]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands

ABSTRACT

In observational studies, treatment assignment is a nonrandom process and treatment groups may not be comparable in their baseline characteristics, a phenomenon known as confounding. Propensity score (PS) methods can be used to achieve comparability of treated and nontreated groups in terms of their observed covariates and, as such, control for confounding in estimating treatment effects. In this article, we provide a step-by-step guidance on how to use PS methods. For illustrative purposes, we used simulated data based on an observational study of the relation between oral nutritional supplementation and hospital length of stay. We focused on the key aspects of PS analysis, including covariate selection, PS estimation, covariate balance assessment, treatment effect estimation, and reporting. PS matching, stratification, covariate adjustment, and weighting are discussed. R codes and example data are provided to show the different steps in a PS analysis. *Am J Clin Nutr* 2016;104:247–58.

Keywords: propensity score, confounding, balance, matching, model selection

INTRODUCTION

Randomized controlled trials (RCTs)[7] are considered the gold-standard approach for estimating the effects of treatments on outcomes. By design, the random assignment of patients to treatment groups ensures that the groups are comparable in both measured and unmeasured baseline characteristics; hence, the association between treatment and outcome is not biased (1–3). As a result, the effect of treatment on outcomes can be estimated by direct comparison of outcomes between treated and untreated groups (3). When RCTs are not feasible for reasons such as cost, time, and ethical issues, the effect of a particular treatment on a certain outcome could be investigated by using a nonexperimental (i.e., nonrandomized) study design. However, in observational studies, treatment selection is influenced by patient baseline characteristics. In the absence of random treatment assignment, systematic differences in baseline characteristics between treatment groups may exist, leading to noncomparability between the groups, which is known as confounding bias. For example, in an observational study of the impact of oral nutritional supplementation (ONS) on hospital length of stay (LOS) (4), patients who received ONS differed from patients who did not receive ONS in their baseline characteristics. Notably, ONS was more often administered to individuals who were less healthy, and hence ONS use could be spuriously associated with increased LOS unless these differences are controlled for. In their seminal article in 1983, Rosenbaum and Rubin (5) introduced propensity score (PS) methods and showed that they can be used to design observational studies and thereby controlling for confounding.

Although PS analysis is a powerful approach and is increasingly being used in observational research (6, 7), errors in the design, analysis, interpretation, and reporting are unfortunately all too common (8). This seems to be in part due to investigators' misunderstanding of the key aspects of the PS methods when conducting and communicating PS analysis (7–9). This article aims to introduce PS analysis, discuss its strengths and limitations, and highlight important steps in the design, analysis, and reporting of PS-based studies. Throughout this article, the different stages of PS analysis are shown by using a hypothetical study based on the observational study on the impact of ONS on LOS (4).

[1] The PROTECT project has received support from the Innovative Medicine Initiative Joint Undertaking (IMI JU; www.imi.europa.eu) under grant 115004, the resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and European Federation of Pharmaceutical Industries and Association (EFPIA) companies' in-kind contributions. In the context of the IMI JU, the Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht University, also received a direct financial contribution from Pfizer.

[2] The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium), which is a public-private partnership coordinated by the European Medicines Agency.

[3] Supplemental Material is available from the "Online Supporting Material" link in the online posting of the article and from the same link in the online table of contents at http://ajcn.nutrition.org.

*To whom correspondence should be addressed. E-mail: o.h.klungel@uu.nl.

[7] Abbreviations used: ATE, average treatment effect; ATT, average treatment effect in the treated; IPTW, inverse probability of treatment weighting; LOS, hospital length of stay; ONS, oral nutritional supplementation; PS, propensity score: RCT, randomized controlled trial.

Received November 4, 2015. Accepted for publication May 26, 2016.
First published online July 13, 2016; doi: 10.3945/ajcn.115.125914.

Supplemental Material can be found at:
http://ajcn.nutrition.org/content/suppl/2016/07/13/ajcn.115.125914.DCSupplemental.html

EXAMPLE DATA SET

To show the different stages of PS analysis, a hypothetical data set was created on the basis of a study on the impact of ONS on hospital outcomes (4). This data set is available in the **Supplemental Material**.

Data sets of 44,000 individuals were created. In line with the motivating example, these data consisted of 11 covariates (X_1–X_{11}), of which 8 were binary covariates (representing sex, congestive heart failure, myocardial infarction, admitted previous 6 mo, admitted from emergency department, diabetes with complications, cancer, and renal disease), 2 were continuous covariates (representing a normally distributed covariate age and a γ distributed covariate the Charlson comorbidity index score), and 1 was a categorical variable [representing race (black, Hispanic, and white)] (4). A binary treatment variable [representing ONS (yes or no)] was then generated by using a logistic regression model, in which treatment status depended on the covariates (X_1–X_{11}), and the coefficients for the covariates for the model were derived from the baseline table in the motivating study (4). We assumed that 1.6% of the simulated patient population received ONS, as in the motivating example (4) and a linear regression model was used to generate outcome data (LOS), conditional on the binary treatment status and the 11 covariates. In the motivating study (4), ONS use was associated with a 2.32-d decrease in LOS (mean difference: −2.3 d). Therefore, we assumed that the mean LOS was 8.42 d for ONS use compared with 10.31 d for non-ONS use. **Table 1** shows the characteristics of the simulated population and the absolute standardized difference in means (proportions) of the characteristics between treatment groups (ONS use compared with non-ONS use).

WHAT IS A PS?

The PS exists in both RCTs and observational studies. In RCTs, the PS is defined by the study design and is known (2). For example, in a simple randomized experiment in which patients are assigned to a treatment or a control group by flipping a coin (assuming equal sample sizes in both groups), the PS for a subject is the probability of being assigned to the treatment group, which is 0.5. In contrast, treatment selection in observational studies is determined by baseline characteristics of the patient and hence the true PS is unknown (2, 5), although it can be estimated by using data on the baseline characteristics of the patient.

Let Z be an indicator variable, denoting the treatment received (ONS episode = 1 and non-ONS episode = 0) and $\mathbf{X}_i$ denotes a vector of baseline characteristics (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, and X_{11}). The PS for patient i (e_i) is the conditional probability (between 0 and 1) of receiving the treatment (ONS = 1 compared with ONS = 0) given the baseline characteristics (X_1–X_{11}): $e_i = \Pr(Z_i = 1|\mathbf{X}_i)$ (5). Intuitively, the PS is a measure of the likelihood that a patient received the treatment (ONS) conditional on his or her covariate values. Hence, it is the summary of all of the covariates included in the PS model and, as such, it has 3 important features. First, it is a balancing score, meaning that at each value of the PS the distribution of the covariates (X_1–X_{11}) defining the PS is expected to be similar in the treated (ONS episode) and untreated groups (non-ONS episode) (5, 10). Second, if treatment assignment is independent of potential outcomes, given the observed covariates (X_1–X_{11}), treatment assignment is also independent of the potential outcomes given the PS (5, 10). Third, the PS needs to be estimated for each patient by using the data, even if the actual treatment status (ONS compared with non-ONS episode) is known (11). Even if the mechanism of treatment assignment is fully known (as in an RCT), the estimated PS performs better than the true PS. This is due to the fact that an estimated PS removes random imbalances in covariates between treatment groups in addition to observed systematic imbalances, whereas the true PS removes only systematic imbalances (12).

The PS should be estimated from the data—for instance, by using logistic regression of the binary treatment (ONS episode compared with non-ONS episode) on the measured covariates (X_1–X_{11}). Applied researchers who have not used PS methods before may question why one should estimate the probability that a patient receives a certain treatment (ONS) although the data clearly show whether a patient has received the treatment (ONS episode). A brief answer to this question is as follows: to create a quasi-randomized experiment by using the patient's

TABLE 1
Baseline characteristics of the hypothetical data[1]

Characteristics	ONS episodes (n = 675)	Non-ONS episodes (n = 43,325)	Absolute standardized difference, %
X_1 = Female, %	53.78	61.45	15.37
X_2 = Congestive heart failure, %	25.33	14.01	26.02
X_3 = Myocardial infarction, %	12.15	8.10	12.37
X_4 = Admitted previous 6 mo, %	41.78	25.74	32.50
X_5 = Admitted from ED, %	58.07	47.36	21.71
X_6 = Diabetes with complications, %	4.44	3.60	4.09
X_7 = Cancer, %	6.96	3.31	14.32
X_8 = Renal disease, %	12.44	8.26	12.66
X_9 = Age, y	65.73	58.82	107.13
X_{10} = Mean Charlson comorbidity index score[2]	3.81	2.30	41.16
X_{11} = Race, %			
Black	17.19	15.64	4.17
Hispanic	6.37	7.34	3.96
White	76.44	77.02	1.36

[1]ED, emergency department; ONS, oral nutritional supplementation.
[2]Charlson comorbidity index score was assumed to have a γ distribution with shape parameter 1 and scale parameter 0.5.

probability of receiving the treatment (i.e., the PS) as a summary score of all measured potential confounders (X_1–X_{11}) and to enable appropriate adjustment of the estimate of the treatment effect (13). This explains one of the key properties of the PS mentioned earlier: if we find 2 patients with the same PS, one in the treated group and one in the untreated group, we can imagine that these 2 patients are more or less "randomly assigned" to one of the treatment groups in the sense of being equally likely to be treated or not. Note that although randomization ensures balance of both measured and unmeasured covariates, PS balances only measured covariates included in the PS model (5, 13, 14). Once the PS is estimated, the PS can be used in 4 different ways to control for confounding; we will come back to that later.

IMPORTANT STEPS IN DESIGNING PS-BASED STUDIES

In the next paragraphs, we outline a step-by-step procedure on how to conduct a PS-based study. Proper PS analysis involves the following steps in sequence: selection of covariates for the PS model, estimation of the PS, choosing one of the PS methods [matching, stratification, covariate adjustment, and inverse probability of treatment weighting (IPTW) using the PS], assessment of covariate balance by using balance metrics, estimation of the treatment effect by using the chosen PS method, and reporting the results. The R code for each step is provided in the Supplemental Material.

COVARIATE SELECTION FOR THE PS MODEL

In many practical settings, investigators encounter high-dimensional data (i.e., large numbers of covariates) with a common exposure and relatively few outcome events. In an attempt to estimate an unbiased causal treatment effect, the selection of important covariates should be made before or during model fitting to avoid problems such as overfitting, particularly when conventional regression methods are being used (15, 16). In such settings, PS methods are invaluable tools for reducing the number of covariates by summarizing the covariate information into a single covariate, the PS.

The selection of covariates can be based on previous knowledge on the relations underlying the data at hand. Different types of variables can be distinguished on the basis of their relation with the treatment (*A*), outcome (*Y*), and other variables: confounding variables [variables that determine treatment status and are also related to the outcome ("X" in **Figure 1**A)], instrumental variables [variables that are strongly related to treatment status but are not related to the outcome, other than through their relation with the treatment ("IV" in Figure 1B); in the motivating example, the fraction of episodes involving any ONS use in a given hospital in a given quarter is used as an instrumental variable], risk factors [variables that are related to the outcome but may not related to treatment status ("R" in Figure 1C)], intermediate variables [posttreatment variables that are influenced by the treatment and lie in the causal pathway from the treatment to the outcome ("I" in Figure 1D)], and colliders [variables that are common effects of 2 causes ("C" in Figure 1E)].

In general, confounding variables and risk factors of the outcome should be included in the PS or regression model to reduce confounding and improve precision of causal effect estimates (17–23). However, adjustment for other types of variables (intermediates, colliders, and instrumental variables) is unnecessary and may even induce bias (17, 20, 23–26).

Despite the popularity of the PS methods, there are no well-developed tools for variable selection in PS models and, as a consequence, applied researchers often use methods that were developed for conventional regression models, such as goodness-of-fit tests (8, 18). However, previous studies showed that such techniques failed to detect variables that should not be adjusted for, such as colliders, intermediate variables, and instrumental variables.

In our example data, we generated all of the covariates in such a way that they are related to both the treatment and the outcome on the basis of the empirical study (4). We did not have instrumental variables, intermediates, and colliders; hence, all of the covariates were selected for inclusion in the PS model.

PS ESTIMATION

Once variables are selected for the PS model, the PS can be estimated by using ordinary logistic regression (although several data-mining techniques, such as neural networks, classification trees, meta-classifiers, and support vector machines, have also been suggested) (13, 27–29). Logistic regression has several advantages. It is a familiar and well-understood statistical tool for investigators and is easy to implement by using standard statistical software packages (29). The inclusion of carefully chosen interactions and square terms in the logistic regression models may improve the balance of covariates in the PS model and reduce the bias in the estimated treatment effect (16, 17).

In the example, we included all of the covariates (X_1–X_{11}), age-squared, and an interaction between age and Charlson comorbidity index score in the PS model (although our data-generating model did not include any interactions or higher-order terms). Hence, a logistic regression model with ONS as the dependent variable and covariates X_1–X_{11} and square and interaction terms as independent variables was fitted to the data. Age-squared and the interaction between age and Charlson comorbidity index score were included in the PS model to improve the balance on age. The predicted probabilities from the model represent the PSs for the patients. The means (ranges) of the PS were 0.071 (0.0014, 0.623) for ONS episodes and 0.014 (0.00003, 0.625) for non-ONS episodes. Logically, the mean probability of receiving ONS (i.e., the mean PS) is larger for ONS episodes than for non-ONS episodes.

PS METHODS TO CONTROL FOR CONFOUNDING

Once the PS is estimated, the next 2 critical steps are to assess the quality of the PS model (see the next section for details how to check this) and to use the PS to actually control for confounding by *1*) creating a matched sample of treated and untreated patients with similar PSs, *2*) stratifying patients on their PSs and estimating treatment effects within the PS strata, *3*) covariate adjustment by using the PS as a covariate, or *4*) IPTW by using the PS (5, 30, 31). These methods are discussed in detail in the section "Estimation and Interpretation of Treatment Effects." At this stage, it might seem puzzling why choosing the type of PS method to be used preceded the assessment of covariate balance. The reason is that the choice of the PS method, which depends on the research question in mind (i.e., the inferential goal of the research), determines how balance of covariates or correct

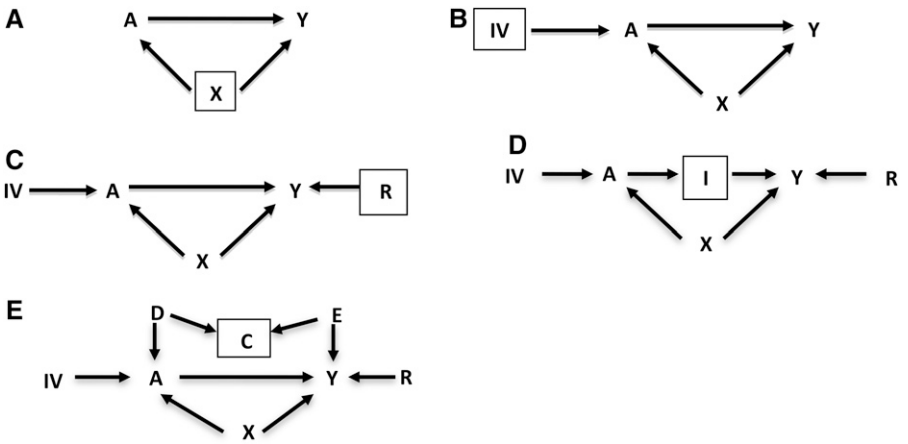

FIGURE 1 Causal diagrams (A–E) depicting different associations between treatment ("A"), outcome ("Y"), confounding variables ("X"), risk factors of the outcome ("R"), instrumental variables ("IV"), intermediate variables ("I"), and common effects that are also called colliders ("C").

specification of the PS model should be evaluated (8, 9, 32). It also dictates the estimation and interpretation of the treatment effect (we will come back to this later) (8, 9, 32).

Here, we considered PS matching, because it is one of the most commonly applied PS methods in the clinical literature. It is also the PS method used in the motivating example. We matched treated and untreated patients using a caliper width of 0.25 on the logit of the PS, meaning that patients were deemed similar if their PS was within a range of 0.25 (on the logit scale) of the PS. The 0.25 caliper width on the logit of the PS is a commonly used caliper in the medical literature (6–8), although the choice of caliper depends on the data at hand and involves a trade-off between precision and bias in the treatment effect estimate. Matching was performed without replacement, although this may increase the bias when a substantial number of untreated subjects are excluded from the analysis (33). In a sensitivity analysis, we used matching with replacement by using a caliper width of 0.25 and matching without replacement by using a caliper width of 0.10; the results were compared with matching by using a caliper width of 0.25 without replacement (Supplemental Material). For matching, we used the MatchIt package in R (33, 34).

ASSESSMENT OF BALANCE ACHIEVED BY THE PS MODEL

The aim of a PS method is to control for confounding by balancing covariates between treatment groups. Therefore, the PS model should be assessed on the basis of its performance in creating balance on covariates and not on how well the PS model discriminates between treated and untreated patients—that is, whether the treatment process is correctly modeled (15) or whether the eventual treatment effect estimates are larger or smaller than expected (10, 15, 35, 36). PS model fitting can be an iterative process where the PS model is updated by including different covariates, interactions, or higher-order terms until an acceptable balance on covariates is achieved (10).

In the literature, different balance measures have been proposed. The absolute standardized difference is more robust in terms of covariate distributions and sample size requirements than are other balance metrics, such as overlapping coefficients (8, 17, 37). It is also a well-understood and easy to calculate statistical tool and is therefore recommended for checking and reporting covariate balances in PS methods (8, 17, 37–40). However, the absolute standardized difference has to be calculated for each covariate, square, or interaction term separately, and there is no consensus on how to pool the covariate-specific standardized differences. Nonetheless, the absolute standardized difference averaged over covariates or square or interaction terms performed better in terms of achieving covariate balance (8, 17, 37, 41). The covariate-specific absolute standardized difference helps to identify the variable that is still imbalanced and to modify the PS model with square and interaction terms of the variable to improve its balance. Alternatively, the absolute standardized difference can be used in combination with a post-matching c-statistic to evaluate balance on all covariates simultaneously (8, 41). Although there exists no universal threshold below which the level of imbalance is always acceptable (42), the use of arbitrary cut-offs for balance diagnostics (e.g., $<$10% for the absolute standardized difference) is prevalent in the medical literature (8).

The use of graphical methods, such as quintile-quintile plots, side-by-side (weighted) box plots, plots of standardized differences of means, and empirical density plots for comparing the distribution of continuous baseline covariates, can provide a quick overview of whether balance has improved for individual covariates (39, 40). Importantly, examining the distribution of PS by using histograms or density plots facilitates subjective judgment on whether there is sufficient overlap between the 2 PS distributions, commonly called "the common support." It can also guide the choice of matching algorithms in PS matching (43). For example, when the overlap in the PS is not substantial, meaning that treated and untreated patients are somewhat different, matching with replacement can be a better option because it will be difficult to find sufficient numbers of untreated matches for the treated patients. When the overlap is too limited, investigators should be aware that the data set, no matter how large, could not support any causal conclusion about the effect of the treatment (1, 42, 44).

In our example, there seems to be sufficient overlap, also known as "common support," in the densities of the PSs (**Figure 2**), which indicates that we can proceed with PS analysis to estimate

treatment effect. Note that sufficient overlap in the PS distributions does not mean sufficient balance on individual covariates. The PS density plots before and after matching are plotted in **Figure 3**. The PS density plots are more similar between the treated and the untreated groups after matching than before matching.

Table 2 shows the balance of covariates after matching on the PS. There is substantial improvement in the balance of covariates in terms of the absolute standardized difference, which is 13.29% at most. The commonly used cut-off for the absolute standardized difference to indicate acceptable balance is 10% (i.e., a standardized difference $<10\%$ is considered a good balance) (8, 35). **Figure 4** shows the percentage change in the absolute standardized difference before and after matching graphically.

Hypothesis-testing statistics, goodness-of-fit tests (e.g., Hosmer-Lemeshow goodness-of-fit test), as well as discrimination tests of a model (c-statistics or the area under the receiver operating characteristics curve of the PS model) should not be used to decide whether the PS model is correctly specified (44, 46). Nevertheless, these are commonly reported in PS-based articles (6–8). Hypothesis tests, such as *t* tests, are functions of both balance and power (i.e., sample size) and, unlike the absolute standardized difference, such tests are not property of a particular sample but they refer to a hypothetical "super-population." Hence, they should not be used as stopping rules for maximizing balance (42). In PS matching, for example, nonsignificant *P* value in a matched sample compared with the original unmatched population might be considered as an indicator of "improved balance." However, the nonsignificant *P* value could be due to a reduced power (sample size) as a result of excluding unmatched subjects, whereas the actual balance may improve, remain the same, or even get worse (39, 42). Similarly, goodness-of-fit tests and discrimination tests of the model neither give an indication of whether an important confounding variable has been omitted from the PS model (45, 46) nor are they related to the degree of covariate balance after conditioning on the PS (1). For example, one can improve the c-statistic of a model by including instrumental variables that might, on the other hand, result in amplification of residual bias due to unmeasured confounding (i.e., exacerbating the imbalance of unmeasured confounders) and reduce the overlap in PS distributions, thereby decreasing the precision of the treatment effect estimate (20, 23, 39, 40).

All balance metrics should be calculated in a way that is similar to how the outcome analysis will be conducted: between matched groups when using PS matching, within strata of the PS when using stratification on the PS, and between treated and untreated patients in the weighted population when using IPTW. When regression adjustment using the PS is used, balance could be assessed by using the standardized difference on the logit of the PS or variance ratios of the residuals of the covariate after adjusting for the logit of the PS (47).

Decisions on whether a PS model has improved covariate balance should be made only on the basis of an examination of patient characteristics measured before any consideration of outcome

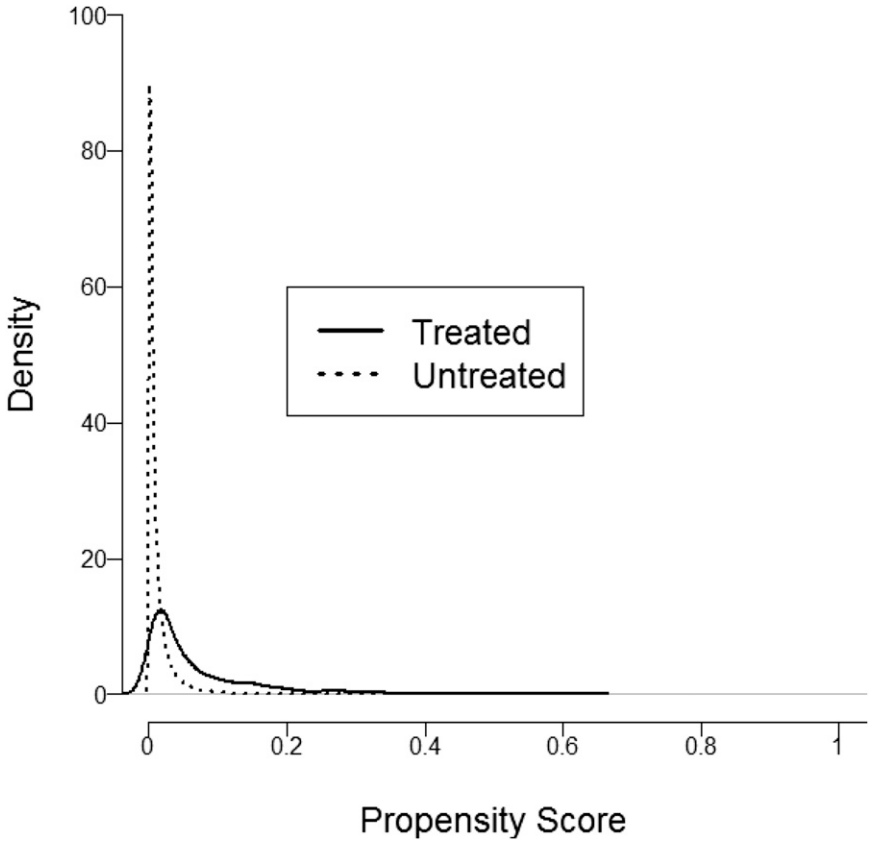

FIGURE 2 Propensity score density plots in treated and untreated patients.

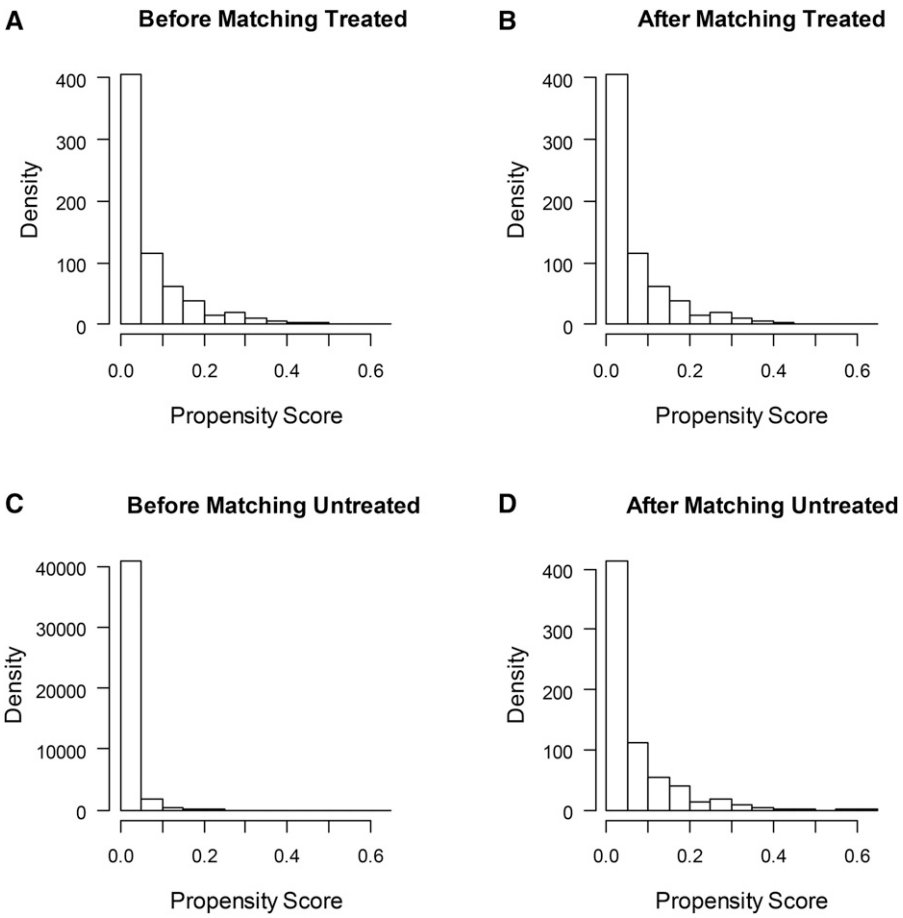

FIGURE 3 Histograms of propensity scores in treated groups before (A) and after (B) matching and in the untreated group before (C) and after (D) matching.

measures (47). Balance metrics evaluate the balance of measured covariates only; they do not indicate the balance of unmeasured covariates. The only way to assess whether unmeasured characteristics are balanced is to collect data on as many characteristics as possible by including "proxies" for unmeasured factors and to examine the balance on measured covariates to which they are related.

ESTIMATION AND INTERPRETATION OF TREATMENT EFFECTS

Different treatment effects can be identified, but the average treatment effect (ATE) in the population and the average treatment effect in the treated subjects (ATT) are of primary interest. ATE refers to the treatment effect if everyone is treated compared with if

TABLE 2
Baseline characteristics of the hypothetical data after matching on the propensity score without replacement[1]

Characteristics	ONS episodes (n = 672)	Non-ONS episodes (n = 672)	Absolute standardized difference, %	Improvement in balance, %
X_1 = Female, %	53.87	52.23	3.29	78.66
X_2 = Congestive heart failure, %	25.30	24.26	2.39	90.80
X_3 = Myocardial infarction, %	12.20	12.80	1.82	85.28
X_4 = Admitted previous 6 mo, %	41.52	42.86	2.71	91.65
X_5 = Admitted from ED, %	58.04	54.32	7.53	65.29
X_6 = Diabetes with complications, %	4.32	4.61	1.44	64.72
X_7 = Cancer, %	6.99	8.04	4.09	71.45
X_8 = Renal disease, %	12.50	12.65	0.45	96.44
X_9 = Age, y	65.66	64.56	13.29	87.59
X_{10} = Mean Charlson comorbidity index score[2]	3.78	3.74	1.31	96.82
X_{11} = Race, %				
Black	17.11	14.58	6.93	−66.19
Hispanic	6.40	8.04	6.70	−69.24
White	76.49	75.45	2.45	−80.20

[1]ED, emergency department; ONS, oral nutritional supplementation.
[2]Charlson comorbidity index score was assumed to have a γ distribution with shape parameter 1 and scale parameter 0.5.

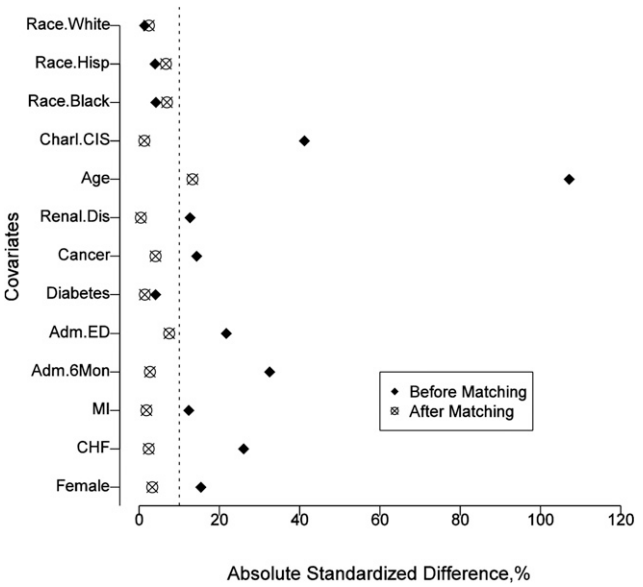

FIGURE 4 Absolute standardized difference plot before matching (all data) and after matching (matched data) for the covariates listed on the *y* axis. The dotted vertical line indicates a commonly used cutoff for absolute standardized difference (10%), which means that a covariate balance <10% in absolute standardized difference is considered acceptable. Adm.ED, admitted from emergency department; Adm.6Mon, admitted previous 6 mo; Charl.CIS, Charlson comorbidity index score; CHF, congestive heart failure; Diabetes, diabetes with complications; MI, myocardial infarction; Race.Hisp, Hispanic; Renal.Dis, renal disease.

everyone is untreated, and ATT refers to the average gain from treatment of those who were actually treated (32).

Matching by using the PS

Once PS matching creates an acceptable balance on important covariates and interactions or square terms by design, causal treatment effects can be estimated by direct comparison of outcomes between matched groups by using differences in means or proportions without the need to rely on parametric models (3, 47). In this case, the analysis is similar to that of an RCT.

PS matching often focuses on estimating the ATT (32), not the ATE, because the closest untreated matches are selected for treated patients, and unmatched patients are often excluded from the analysis. When there is limited overlap in the PS distribution between treatment groups, treated and untreated patients in the extremes of the PS distributions (i.e., the nonoverlapping regions) should be excluded because one cannot infer treatment effects in this region without extensive extrapolation. It is important to note that the exclusion of unmatched patients 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. For example, the exclusion of untreated subjects due to a lack of close matches changes the estimate from ATT to the ATE in the treated patients for whom we can find untreated matches (32). One can also estimate the ATE in the sample with modifications of the matching algorithms. For example, full matching, which uses all patients for analysis, can estimate either the ATT or the ATE (35, 36). Although matching, in general, discards some data (i.e., unmatched patients), it can actually increase the efficiency of treatment effect estimates (48, 49).

In our example data, we used 1:1 nearest-neighbor matching without replacement; the matched data comprised treated patients for whom untreated matches were found and their untreated matches. Among a patient population of 44,000 (non-ONS use = 43,325, ONS use = 675), 672 of the treated patients were matched to 672 untreated patients. For 3 of the treated patients, there were no untreated matches; hence, the treatment effect we estimated was the ATT for whom we found untreated matches. The unadjusted analysis resulted in a mean difference of −1.89 d (95% CI: −1.97, −1.80 d), which was biased compared with the PS-matched analysis, and ONS use lowered the LOS by 2.38 d (95% CI: −2.50, −2.26 d) (**Table 3**).

Stratification using the PS

Within the strata formed by the PS, measured covariates are assumed to be balanced between treatment groups; hence, the treatment effect can be estimated by direct comparison of outcomes between treated and untreated patients (2, 14). The stratum-specific treatment effects can then be aggregated across subclasses to obtain an overall measure of treatment effect (2). It 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 subjects in each stratum provides ATT, whereas weighting by the total number of subjects in each stratum yields the ATE (50). 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 there is a homogeneous treatment effect among the PS strata. When there is heterogeneity of treatment effect among the PS strata, pooling the stratum-specific treatment effect complicates the interpretation of the treatment effect. Alternatively, the quintiles and deciles of the PS can be used as a categorical variable in a model-based adjustment to estimate treatment effects (10). Similar to PS matching, stratification using the PS relies less on parametric models. By using model-based adjustment on the quintiles and

TABLE 3
Association between ONS use and hospital length of stay by using linear regression and different PS methods[1]

Methods	Mean difference (95% CI)[2]
Crude	−1.89 (−1.97, −1.80)
Conventional linear regression	−2.41 (−2.49, −2.33)
PS matching	−2.38 (−2.50, −2.26)
Stratification[3]	
Quintiles of PS	−2.29 (−2.37, −2.21)
Deciles of PS	−2.34 (−2.42, −2.26)
Covariate adjustment using PS	−2.40 (−2.49, −2.32)
IPTW	
Unstabilized	−2.37 (−2.50, −2.24)
Stabilized	−2.37 (−2.50, −2.24)

[1]True mean difference = −2.32. IPTW, inverse probability of treatment weighting; ONS, oral nutritional supplementation; PS, propensity score.
[2]Mean difference is the difference in the hospital length of stay in days.
[3]Model-based adjustment by using quintiles and deciles of the PS as a categorical variable.

deciles of the PS (i.e., using the quintiles and deciles of the PS as a categorical variable) in the regression model, ONS use lowered the LOS by 2.29 d (95% CI: −2.37, −2.21d) and 2.34 d (95% CI: −2.42, −2.26 d), respectively.

Regression adjustment using the PS

To control for confounding, one can also include the PS, in addition to the treatment variable, as a covariate in a regression model (i.e., the outcome variable is regressed on the treatment variable and the estimated PS). Although covariate adjustment using the PS is easy to apply, it is considered to be a suboptimal application of the PS for ≥3 reasons. First, treatment effect estimation is highly model-dependent because it mixes the study design and data analysis steps, hence it requires correct specification of the PS model (15, 16). Second, it makes additional assumptions unique to regression adjustment, namely that the relation between the outcome and the estimated PS must be linear and that there should be no interaction between treatment and the PS (2, 5, 47). Third, although it generally allows estimation of the ATE, the interpretation is complicated in the case of noncontinuous outcomes where the estimate of interest is noncollapsible. Noncollapsibility refers to a phenomenon in which, in the presence of a non-null treatment effect, the (overall) marginal treatment effect estimate is different from the (stratum-specific) conditional treatment effect estimate, even in the absence of confounding (e.g., OR or HR) (9, 24, 25).

When the PS is used as a covariate in a regression model, the ATE and ONS use was associated with a 2.40-d decrease in LOS (95% CI: −2.49, −2.32 d). There was no significant interaction between treatment and the PS.

IPTW by using the PS

IPTW, like PS matching and stratification, can be viewed as a method involving pre-processing the data by using weights to create an "artificial" population, called a "pseudo-population," in which treatment is independent of measured covariates (30, 31). As a consequence, one can estimate the treatment effect by direct comparison of outcomes between treated and untreated patients. Alternatively, the weights can be used in weighted regression models to estimate the treatment effect. Although this method focuses on estimating the average effect in the population (ATE), modification of the weights allows us to estimate the ATT (51). Most important, 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 (52), or bootstrapping. The downside of this approach is that when some patients have probabilities receiving the treatment close to 0 or 1, the weights for such patients become unstable. To address this problem, stabilizing the weights has been proposed to "normalize" the range of the inverse probabilities and to increase the efficiency of the analysis (30, 31, 53).

In IPTW, weights are assigned to treated or untreated patients as the inverse of the probability of receiving their own treatment: 1/PS for treated patients and 1/(1 – PS) for untreated patients. In the example data, the mean, median, and range of the weights were 1.89, 1.01, and 1.0–706.3, respectively, without stabilization. The IPTWs were stabilized by replacing the "1" in the numerator of the weight by the proportion of ONS and non-ONS episodes in the treated and untreated populations, respectively. Accordingly, the mean, median, and range of the stabilized weights were 1.0, 0.99, and 0.02–10.8, respectively. In this particular example, the weight stabilization did not affect the treatment effect. ONS was associated with a 2.37-d (95% CI: −2.50, −2.24 d) decrease in LOS by using both unstabilized and stabilized IPTW.

All of the above PS methods have their own advantages and limitations. **Table 4** describes and compares the 4 different PS methods in terms of their use at the design or analysis stage, covariate balance assessment, model dependence, and the treatment effect they can estimate and its interpretation.

REPORTING OF PS ANALYSIS

PS methods are invaluable tools in observational studies. However, like regression analysis, the quality of the results obtained from PS analysis depends on appropriate conduct using the consecutive steps. For a critical appraisal of a PS-based study, the reader has to rely on the information provided. Despite substantial developments and common applications of PS methods, reporting of aspects of the PS analysis is generally poor and inconsistent in the medical literature (7, 8, 54). This could be, in part, due to a lack of standards for conduct as well as reporting of PS methods in guidelines on the reporting of observational studies, such as the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement (55, 56). Details on important aspects of PS analysis that should be reported are included in **Figure 5** (8).

STRENGTHS AND LIMITATIONS OF PS METHODS

PS methods are primarily aimed balancing treatment groups with respect to covariate distributions; when such balance is achieved, it is relatively easy to detect and communicate (15) by using simple statistics or plots. Similarly, PS methods, unlike regression methods, can also warn investigators that, due to inadequate overlap in covariate distributions (i.e., poor "common support") between treatment groups, a particular data set cannot address the causal question without relying on untrustworthy

TABLE 4
Comparison of the different PS methods and their advantages and disadvantages[1]

Method	Description	Advantages	Disadvantages
PS matching	Constructs treated and untreated matched groups with similar PSs Primarily estimates ATT, but ATE can also be estimated with slight modifications	Straightforward and easy to apply Minimizes model dependence Separates the study design and data analysis stage of a study Easy to check improvements on covariate balance	No consensus on variance estimations Interpretations can be complicated, particularly when some observations are excluded
PS stratification	Constructs strata of treated and untreated subjects with closer PSs Can estimate ATT or ATE	Straightforward and easy to apply Separates the study design and data analysis stage of a study Minimizes model dependence Easy to check improvements on covariate balance	Interpretations can be complicated, particularly in the presence of treatment effect modification by the PS Residual confounding depending on the number of strata used Balance assessment can be laborious compared with PS matching
Regression adjustment using PS	PS is used as a single summary of all covariates (included in PS model) in regression model Estimates ATE	Straightforward and easy to apply even compared with all other PS methods	Checking improvements on covariate balance is not straightforward Requires correct specification of PS model Mixes up the design and analysis stages of a study and focuses more on the analysis stage than the design stage Relies on the assumption of linear relations between the PS and outcome Interpretations could be complicated when noncollapsible effect measures such as ORs are used Extrapolates even when there is no positivity[2]
Inverse probability of treatment weighting using PS	PSs are used as weights to create a pseudo-population in which exposure becomes independent of measured covariates included in the treatment (PS) model Can estimate ATT or ATE	Easy to apply Extends to time-varying treatment and confounding setting	Focuses more on the analysis stage than the design stage Requires correct specification of PS and outcome model Sensitive to observations with extreme weights and nonpositivity Slightly complicated compared with stratification and regression adjustment using PS

[1]ATE, average treatment effect; ATT, average treatment effect in the treated patients; PS, propensity score.
[2]Positivity requires that there be both treated and untreated patients at every combination of the values of the measured confounder or confounders in the population under study.

"model-dependent" extrapolations (15, 16, 43). The investigator might opt for restricting the conclusion or the inference to the group of patients sufficiently represented in both treatment groups by using methods such as nearest-neighbor matching with caliper widths that will result in excluding patients in the nonoverlapping regions of the PS. In this case, the treatment effect estimate will be the ATT for whom we found untreated matches.

Like randomized experiments, PS matching allows for designing a study separate from the analysis part of the study (i.e., first, covariate balance can be achieved by using PS matching without using the outcome variable, and then treatment effect is estimated in the matched data without relying much on assumptions underlying the outcome model, such as correct model misspecification). Hence, causal inference can be made with minimal model-dependence (5, 15, 16, 57). It is important to note that if nonparametric preprocessing of the data using PS matching results in no reduction in 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 (15, 57).

PS methods also provide an efficient way to control for covariates or potential confounding variables when the number of outcome events is limited compared with the number of covariates, thereby minimizing the "curse of high dimensionality" in the data because the PS as a single covariate summarizes the covariates included in the PS model (5, 10, 43, 57). In fact, previous studies suggest that ≥10 outcome events are required for every covariate included in a regression model (58–60). Hence, reducing a large number of confounders into a single PS can be beneficial in case of a limited number of observations available for analysis.

Given the widespread applications of PS methods for addressing causal questions using observational studies, it is crucial to keep in mind that PS methods, like other regression methods, can only control for measured confounding variables and not for unmeasured ones (5, 10). As a result, PS analysis can only be as good as the quality and the completeness of potential confounding variables that are at the disposal of the researcher. Only a rich set of covariates may convince a critical reader that no unmeasured confounding variables were missed. Therefore, it is important that investigators provide a detailed account of

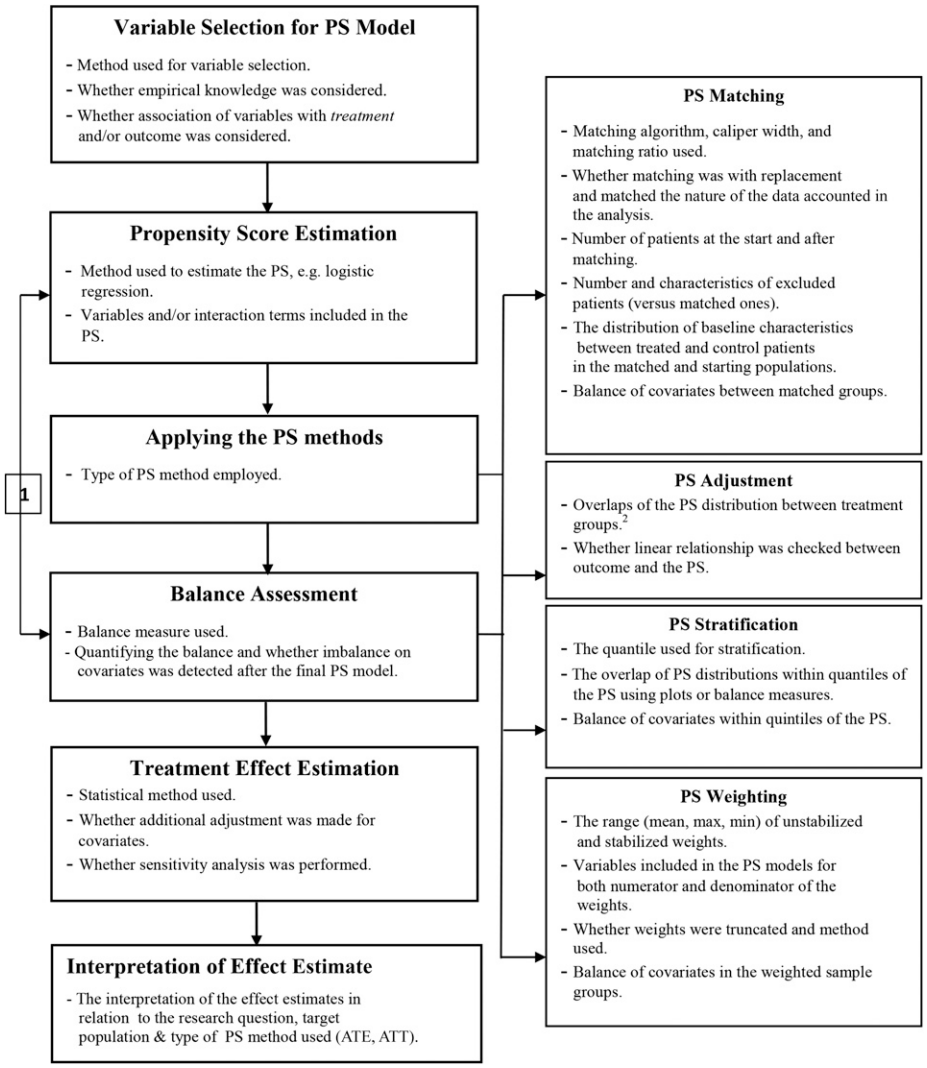

FIGURE 5 Flowchart summarizing relevant information that should be reported when conducting a PS analysis. [1]The PS estimation can be iteratively continued until an "optimal" balance on covariates is reached. [2]It is not relevant to report goodness-of-fit tests, prematching c-statistics of the PS model, the actual PS values, the PS model itself, or *P* values and model coefficients from the PS model. ATE, average treatment effect; ATT, average treatment effect in the treated patients; max, maximum; min, minimum; PS, propensity score.

the variables collected and included in the PS model. Furthermore, sensitivity analyses (5, 11, 48) are invaluable tools to assess the plausibility of the assumptions underlying the PS methods and how violations of them might affect the conclusions (61).

An additional limitation of PS methods is that they work better in large samples (43) because the distributional balance achieved on measured covariates is an expected balance. As a result, in smaller studies, an imbalance of covariates is inevitable even if the PS model is correctly specified, whichever PS method is used. As a consequence, investigators attempting to answer causal questions with the use of observational studies should explore large data sets with reasonable qualities.

CONCLUSIONS

In conclusion, PS methods are invaluable tools for estimating treatment effects from observational data in a transparent way. They should neither be regarded as a "panacea for the deficiencies of observational studies nor as replacement for model-based adjustments, but as critical tools contributing to their initial designs" (15), and they could be used in combination with model-based adjustment methods to minimize model-dependence. Taking full advantage of the methods requires, in addition to the initial study design, the detailed specification of all statistical analyses to be performed. In addition, adequate reporting of different aspects of the PS analysis is as crucial as the analysis itself because readers depend on the information reported to judge the quality of the analysis and validity of the results, as do other investigators who would want to replicate the study.

The authors' responsibilities were as follows—MSA conducted the analysis of the data and wrote the first draft of the manuscript; RHHG and OHK critically reviewed the draft version of the manuscript; and all authors: contributed to the concept and design of the study and the interpretation of the results of the simulated data set, and read and approved the final version of the manuscript. The views expressed are those of the authors only and not of their respective institution or company. OHK received unrestricted funding for pharmacoepidemiologic research from the Dutch private-public

funded Top Institute Pharma; there were no financial, personal, political, academic or other relations that could lead to a conflict of interest. The remaining authors did not declare any conflicts of interest.

REFERENCES

1. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med 2007;26:734–53.
2. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46:399–424.
3. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66:688–701.
4. Philipson TJ, Thornton Snider J, Lakdawalla DN, Stryckman B, Goldman DP. Impact of oral nutritional supplementation on hospital outcomes. Am J Manag Care 2013;19:121–8.
5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
6. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. J Thorac Cardiovasc Surg 2007;134:1128–35.
7. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008;27:2037–49.
8. Ali MS, Groenwold RH, Belitser SV, Pestman WR, Hoes AW, Roes KC, de Boer A, Klungel OH. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol 2015;68:112–21.
9. Ali MS, Groenwold RHH, Klungel OH. Propensity score methods and unobserved covariate imbalance: comments on "squeezing the balloon". Health Serv Res 2014;49:1074–82.
10. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. JASA 1984;79:516–24.
11. Rosenbaum PR. Model-based direct adjustment. JASA 1987;82:387–94.
12. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol 1999;150:327–33.
13. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–81.
14. D'Agostino RB Jr. Propensity scores in cardiovascular research. Circulation 2007;115:2340–3.
15. Rubin DB. On principles for modeling propensity scores in medical research. Pharmacoepidemiol Drug Saf 2004;13:855–7.
16. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007;26:20–36.
17. Ali MS, Groenwold RHH, Pestman WR, Belitser SV, Roes KCB, Hoes AW, Boer A, Klungel OH. Propensity score balance measures in pharmacoepidemiology: a simulation study. Pharmacoepidemiol Drug Saf [Internet]. 2014;23:802–11.
18. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149–56.
19. Patrick AR, Schneeweiss S, Brookhart MA, Glynn RJ, Rothman KJ, Avorn J, Stürmer T. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. Pharmacoepidemiol Drug Saf 2011;20:551–9.
20. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol 2011;174:1213–22.
21. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Glynn RJ. Myers et al. respond to "Understanding bias amplification". Am J Epidemiol 2011;174:1228–9.
22. Pearl J. On a class of bias-amplifying variables that endanger effect estimates. In: Grunwald P and Spirtes P, editors. Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010). Corvallis, Oregon: Association for Uncertainty in Artificial Intelligence; 2010;417–24.
23. Pearl J. Invited commentary: understanding bias amplification. Am J Epidemiol 2011;174:1223–7.
24. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology 2003;14:300–6.
25. Greenland S, Pearl J. Adjustments and their consequences—collapsibility analysis using graphical models. Int Stat Rev 2011;79:401–26.
26. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Stat Sci 1999;14:29–46.
27. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. PLoS One 2011;6:e18174.
28. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf 2008;17:546–55.
29. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J Clin Epidemiol 2010;63:826–33.
30. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–60.
31. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000;11:561–70.
32. Hill J. Discussion of research using propensity-score matching: comments on "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003" by Peter Austin, Statistics in Medicine. Stat Med 2008;27:2055–61.
33. Ho D, Imai K, King G, Stuart E. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 2011;42:1–28.
34. R Core Team. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; 2012. Available from: http://www.R-project.org/.
35. Stuart EA, Green KM. Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. Dev Psychol 2008;44:395–406.
36. Hansen BB. Full matching in an observational study of coaching for the SAT. JASA 2004;99:609–18.
37. Belitser SV, Martens EP, Pestman WR, Groenwold RHH, Boer A, Klungel OH. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf 2011;20:1115–29.
38. Groenwold RHH, Vries F, Boer A, Pestman WR, Rutten FH, Hoes AW, Klungel OH. Balance measures for propensity score methods: a clinical example on beta-agonist use and the risk of myocardial infarction. Pharmacoepidemiol Drug Saf 2011;20:1130–7.
39. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28:3083–107.
40. Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. Pharmacoepidemiol Drug Saf 2008;17:1202–17.
41. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. Stat Med 2014;33(10):1685–99.
42. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. J R Stat Soc Ser A Stat Soc 2008;171:481–502.
43. Dehejia R, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat 2002;84:151–61.
44. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997;127:757–63.
45. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf 2011;20:317–20.
46. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. Pharmacoepidemiol Drug Saf 2005;14:227–38.
47. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Serv Outcomes Res 2001;2:169–88.
48. Rosenbaum PR. Observational study. Encyclopedia of statistics in behavioral science. 2005.
49. Smith HL. Matching with multiple controls to estimate treatment effects in observational studies. Sociol Methodol 1997;27:325–53.
50. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci 2010;25:1–21.
51. Morgan SL, Todd JJ. A diagnostic routine for the detection of consequential heterogeneity of causal effects. Sociol Methodol 2008;38:231–81.
52. Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models. Am Stat 2004;58:272–9.

53. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol 2006;163:262–70.
54. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf 2004;13:841–53.
55. Jadrijević-Mladar Takač M. ENCePP guide on methodological standards in pharmacoepidemiology. 3rd PharmSciFair. Pharmaceutical Sciences for the Future of Medicines; Final Programme and Book of Abstracts/Lådan & Co AB (ur.). Stockholm, Sweden: European Federation for Pharmaceutical Sciences, EUFEPS, 2011. p. 51.
56. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Prev Med 2007;45:247–51.
57. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 2007;15:199–236.
58. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995; 48:1503–10.
59. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.
60. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003;158:280–7.
61. Stuart EA, Rubin DB. Matching methods for causal inference: designing observational studies. In: Osborne J, editor. Thousand Oaks (CA): Sage Publications; 2007. p. 1.