Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners

Rishi J Desai,[1] Jessica M Franklin[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street, Boston, MA 02120, USA

Correspondence to: R J Desai
rdesai@bwh.harvard.edu
(ORCID 0000-0003-0299-7273)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2019;367:l5657
http://dx.doi.org/10.1136/bmj.l5657

Accepted: 5 August 2019

This report aims to provide methodological guidance to help practitioners select the most appropriate weighting method based on propensity scores for their analysis out of many available options (eg, inverse probability treatment weights, standardised mortality ratio weights, fine stratification weights, overlap weights, and matching weights), and outlines recommendations for transparent reporting of studies using weighting based on the propensity scores.

Propensity scores[1] have become a cornerstone of confounding adjustment in observational studies evaluating outcomes of treatment use in routine care. Propensity score based methods target causal inference in observational studies in a manner similar to randomised experiments by facilitating the measurement of differences in outcomes between the treated population and a reference population.[2] Despite the conceptual equivalence between randomised experiments and observational studies using propensity scores, randomised experiments can successfully achieve exchangeability between treated and reference populations with respect to both measured and unmeasured characteristics, whereas observational studies can only achieve exchangeability with respect to the measured characteristics.

Propensity scores, formally defined as patients' predicted probability of receiving a certain treatment given their characteristics, need to be estimated using observed data based on a statistical model such as a logistic regression model. After estimation, confounding adjustment through conditioning on the propensity scores can be done in many ways, including matching, stratification, adjustment as a regressor, and weighting.[3] Previous research has suggested that the traditional outcome regression model provides generally equivalent confounding adjustment to various propensity score based approaches in cohort studies with a large sample size and sufficient number of outcome events to support multivariable model fit.[4] However, some key advantages of propensity scores, including the ability to clearly define the target population of inference and the ability to identify and exclude patients in atypical circumstances with near zero probability of receiving a certain treatment,[5] have made use of these scores a method of choice for analysing observational data for many researchers.

Matching each treated observation to a fixed number of reference observations if their propensity scores are within a prespecified range (the caliper) has often been the preferred approach of using propensity scores for confounding adjustment.[6] However, this method has an important limitation of discarding unmatched observations falling within the caliper after a pre-specified number of observations are found for each treated observation. More recently, a paradoxical phenomenon of increasing rather than decreasing covariate imbalance after propensity score matching has been described by King and Nielsen.[7] Notably, other methods of using propensity scores in analysis (including stratification, adjustment as a regressor, and weighting) are not affected by this paradox.

Weighting on the propensity score has several advantages. Firstly, unlike matching, weighting keeps most observations in the analysis and hence, can offer increased precision when estimating treatment effects. Secondly, unlike regression adjustment by the propensity score,[8] weighting lends itself easily to transparent reporting of the balance achieved between treatment and reference populations. Finally, weighting on the propensity score is arguably the most flexible approach of using propensity scores in the analysis with multiple available variations that allow targeting specific populations for inference. In addition to traditional approaches of propensity score weighting that use inverse probability treatment weights (IPTW) or standardised mortality ratio weights (SMRW), several newer approaches (including propensity score fine stratification weights,[9] matching weights,[10 11] and overlap weights[12]) have been proposed to overcome important limitations of traditional weighting approaches.

SUMMARY POINTS

Propensity score based weighting approaches provide an alternative to propensity score matching and are especially useful when preserving a large majority of the study sample is needed to maximise precision

Propensity score based weighting approaches can target treatment effect estimation in specific populations including the average treatment effect in the whole population, average treatment effect among the treated population, or average treatment effect in a subpopulation with clinical equipoise

Principles outlined in this report are intended to help investigators in identifying the most suitable propensity score based weighting approach for their analysis and provide a framework for transparent reporting

In this report, we describe implementation of alternative propensity score weighting methods along with key features of each approach to help practitioners choose the most appropriate method for their analysis. We also provide recommendations for key diagnostic and reporting parameters to evaluate the validity of an analysis using propensity score weighting. The objective of this report is not to compare performance of different weighting methods, but rather to demonstrate implementation and provide insights into the process of selecting a specific approach for a particular study. For additional technical details and comparative performance of the various weighting approaches described here, we refer readers to previously published studies that have proposed and rigorously evaluated these approaches under various scenarios.[4 5 8-14]

Basic principle of weighting methods based on propensity scores

The propensity score is a balancing score that allows for simultaneous balance on a large set of covariates between the treated and reference populations. Matching and traditional stratification of the propensity score (also referred to as subclassification)[1] achieve balance by ensuring that treated and reference populations on average have comparable propensity scores (within each stratum if using subclassification). However, weighting methods use a function of the propensity score to reweight the populations and achieve balance by creating a pseudo-population where the treatment assignment is independent of the observed covariates.[15] A weighted outcome regression model can be implemented with treatment status as the only independent variable to derive adjusted treatment effect estimates, because covariates are expected to be balanced in the weighted population. To account for the fact that the pseudo-population size is inflated or deflated relative to the original study population and that weights are estimated (rather than known with certainty), a robust, sandwich type estimator is recommended for variance estimation for the treatment effect estimates.[16]

Target of inference (estimand)

The target of inference refers to the patient population to which the estimated treatment effect applies and will generally be study specific. Investigators should consider the following central question when conceptualising the target of inference for a specific study—would it be feasible to treat all eligible patients included in the study with the treatment of interest?

If the answer to this question is yes, then the target of inference might be defined as the average treatment effect (ATE). An example of where ATE could be the target of inference might be in a study comparing the effectiveness of a newly approved treatment with an existing treatment for a certain condition, for example, dabigatran versus warfarin for prevention of stroke in atrial fibrillation.[17] Because both of these treatments are indicated as exchangeable options for atrial fibrillation in the absence of specific contraindications, all patients meeting the study inclusion criteria—namely, the diagnosis of atrial fibrillation—are eligible to receive dabigatran.

If the answer to the central question is no, the treatment would not be given to everyone in the eligible population, and only patients with certain characteristics who actually received the treatment would be ideal candidates for treatment; then the target of inference might be defined as average treatment effect among the treated population (ATT). An example of where ATT could be the target of inference might be in a study evaluating the safety of a particular drug treatment or class in a population of vulnerable patients, for example, antipsychotic drugs for pregnant women.[18] Because of the concerns and uncertainty related to malformation risks associated with antipsychotics, not all patients meeting the study inclusion criteria—namely, the diagnosis of schizophrenia, bipolar disorder, or psychosis—might be considered for treatment. Therefore, only women with greater severity of these conditions would receive treatment with antipsychotics during pregnancy, making the ATT the relevant target of inference. There might also be circumstances when the interest is in targeting ATE only among a subset of patients with certain characteristics leading to clinical equipoise. Weighting approaches based on the propensity score can accommodate all three of these targets of inference. The key features and mathematical formulas of each weighting approach are summarised in table 1 and described in detail below. In the absence of treatment effect heterogeneity by patient characteristics, ATE and ATT will coincide.

Considerations when selecting a propensity score weighting method for confounding adjustment

We describe a stepwise process (fig 1) that investigators can consider when selecting an appropriate weighting method based on the propensity score for their study. We use a cohort study of dabigatran versus warfarin initiation on the risk of ischaemic stroke or systemic embolism conducted using commercial insurance claims data from the United States[19] as a recurring case study throughout this manuscript to demonstrate various concepts as they relate to alternate propensity score weighting approaches.

Step 1: Correct specification of the propensity score model

The first critical step in an analysis using the propensity score for confounding adjustment is avoiding misspecification of the propensity score model. Because an investigator is unlikely to know the true structural association between treatment assignment and all covariates, model misspecification is possible when estimating the propensity score from a simple logistic regression model that only includes main effects and not interactions among variables. Other approaches to estimate the propensity score—for instance, the covariate balancing propensity scores or machine

doi: 10.1136/bmj.l5657 | BMJ 2019;367:l5657 | the bmj

Table 1 | Alternative approaches for weighting based on propensity scores

Method	Weight calculation: Treated patients	Weight calculation: Reference patients	Target of inference (estimand)	Features	Interpretation
Inverse probability of treatment weights	1/PS	1/(1 – PS)	ATE in the whole population	Clear target of inference, which mimics the target of inference from randomised controlled trials, is a strength. However, because the PS is directly used to create weights, extreme weights are commonly observed. Weight trimming is routinely necessary to address extreme weights and prevent variance inflation	ATE estimates can be interpreted as effect of the treatment when the whole study population is treated with the treatment under investigation versus the reference treatment
Fine stratification weights (ATE)	$(N_{\text{total in PS stratum i}}/N_{\text{total}})$ / $(N_{\text{exposed in PS stratum i}}/N_{\text{total exposed}})$	$(N_{\text{total in PS stratum i}}/N_{\text{total}})$ / $(N_{\text{reference in PS stratum i}}/N_{\text{total reference}})$	ATE in the whole population	Does not use the PS directly to calculate weights; instead, the scores are used to create fine stratums and weights are subsequently calculated to account for stratum membership. As a result, extreme weights due to PSs that are very close to 0 or 1 are unlikely: an important strength in circumstances where exposure prevalence is low. Clear target of inference is another strength	
Standardised mortality ratio weighting	1	PS/(1 – PS)	ATT	Weighting is conducted by the odds in the reference group, can naturally extend to circumstances with >2 treatment arms. Weight trimming might be necessary to address extreme weights and prevent variance inflation. Clear target of inference is a strength	ATT estimates can be interpreted as effect of the treatment when patients receiving treatment in the study population (that is, the exposed group) were treated with the treatment under investigation versus the reference treatment
Fine stratification weights (ATT)	1	$(N_{\text{exposed in PS stratum i}}/N_{\text{total exposed}})$ / $(N_{\text{reference in PS stratum i}}/N_{\text{total reference}})$	ATT	Does not use the PS directly to calculate weights; instead, the scores are used to create fine stratums and weights are subsequently calculated to account for stratum membership. As a result, extreme weights due to PSs that are very close to 0 or 1 are unlikely: an important strength in circumstances where exposure prevalence is low. Clear target of inference is another strength	
Matching weights	(Minimum (PS, 1 – PS)) / PS	(Minimum (PS, 1 – PS)) / (1 – PS)	ATE in a subset	Extreme weights are impossible because weights are bound between 0 and 1 by design, eliminating the need for weight trimming. Can naturally extend to circumstances with more than two treatment arms	Target of inference is close to ATE in the whole population when groups are equally sized and PS distributions have good overlap, and is close to the ATT in the smaller group when groups are unequally sized but PS distribution have good overlap. In circumstances of limited overlap in PS distribution, could lead to treatment effect estimation in a subpopulation that does not reflect patients receiving the treatment of interest in routine care or the whole study population
Overlap weights	(1 – PS)	PS	ATE in the overlap population	Extreme weights are impossible because weights are bound between 0 and 1 by design, eliminating the need for weight trimming. Yields exact covariate balance between treated and reference groups by construction	Estimates can be interpreted as ATE when patients with a realistic probability of receiving either treatment were treated with the treatment under investigation versus the reference treatment. The target population in this approach can be described as the overlap population or population with reasonable clinical equipoise for treatment decision. However, this approach could lead to treatment effect estimation in a subpopulation that does not reflect patients receiving the treatment of interest in routine care or the whole study population, especially when PS overlap is limited

ATE=average treatment effect; ATT=average treatment effect among the treated population; PS=propensity score.

learning approaches such as neural networks—could provide alternatives that are less prone to misspecification.[20 21] Regardless of the approach used for constructing propensity score models, researchers should emphasise inclusion of outcome risk factors in the model[22] and exclusion of strong predictors of treatment that are not associated with outcomes (that is, an instrumental variable) from the model to avoid increased variance and amplification of bias due to unmeasured confounding.[23]

When considering weighting based on the propensity score, the impact of model misspecification could vary across approaches. Approaches that use the score directly to create weights such as IPTW

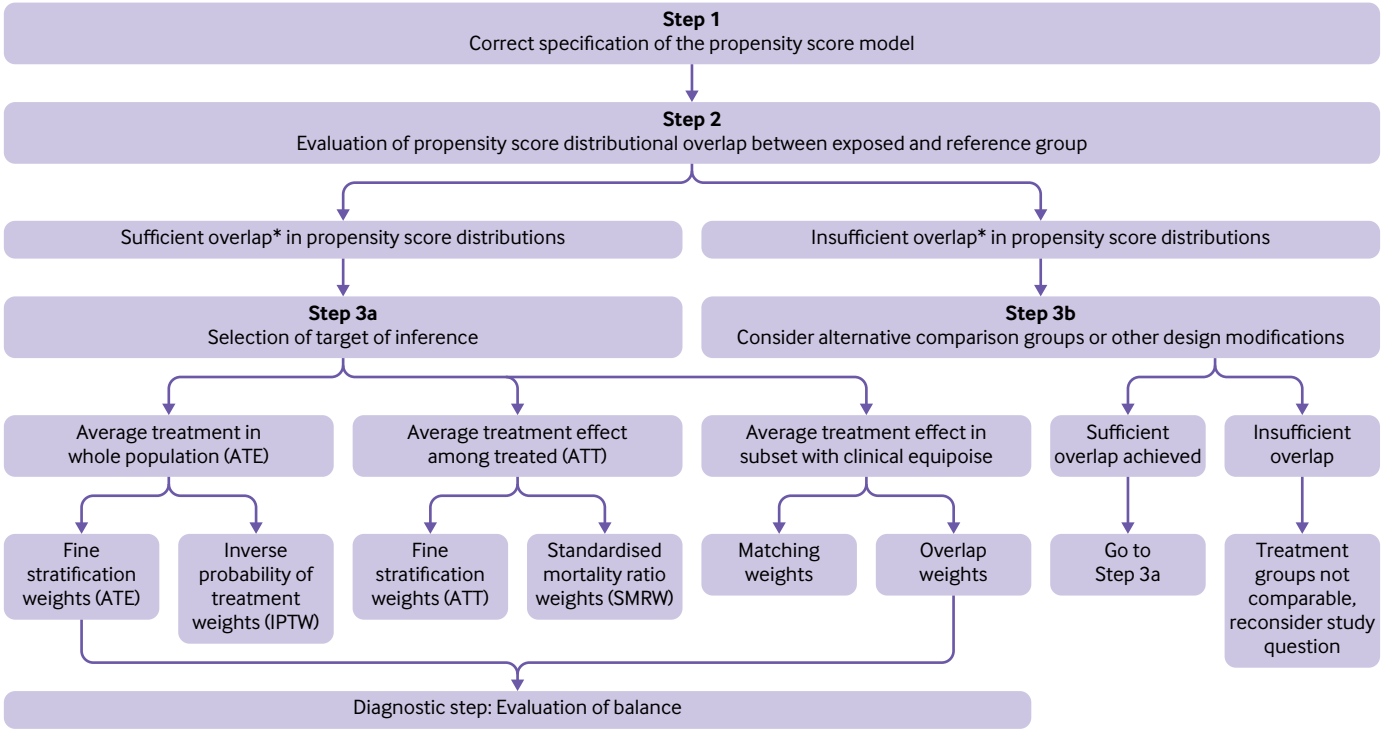

Fig 1 | Factors to consider in the selection of a propensity score weighting method for confounding adjustment. *If a large portion of the sample is lost after trimming non-overlapping regions of propensity score distributions, it might indicate insufficient overlap between distributions.

are theoretically more prone to increased bias and variance from misspecification of the propensity score model.[24 25] On the other hand, the weighting approach based on propensity score stratification might be more robust against misspecification of the propensity score model, because it can be conceptualised as a semiparametric implementation of propensity score weighting that uses the score only to create stratums and then uses the counts of observations within each stratum to derive weights. A simple diagnostic step of checking covariate balance between the treatment and reference populations after applying weights based on the propensity score can alert researchers to potential model misspecification that might need attention.[20] For reporting the balance in each individual covariate between treated and reference populations, a measure such as the standardised difference in prevalence (or means for continuous variables) is recommended.[26] Investigators might also consider reporting an overall measure of balance, such as the post weighting C statistic, where values closer to 0.5 would indicate achievement of balance in aggregate over all included covariates.[27] Box 1 summarises the recommended diagnostic and reporting steps for analyses conducted using propensity score based weighting.

In the case example, a propensity score model was constructed with dabigatran initiation as the dependent variable and 66 prespecified patient characteristics as independent variables in a logistic regression model in a cohort of 79 265 patients with atrial fibrillation, 22 809 (29%) of whom were dabigatran initiators. Conditioning on the propensity scores derived from this model through various weighting approaches (described below) led to balance among included covariates (table 2, fig 2). Achievement of balance suggests that propensity score model specification was probably adequate in this example.

Step 2: Evaluation of propensity score distributional overlap between exposed and reference groups

Evaluation of the propensity score distributional overlap between the treatment and reference groups is the next important step in an analysis using the propensity score. High overlap in the propensity score distribution generally indicates a reasonable degree of clinical equipoise in treatment selection between the comparator groups. The general recommendation of trimming the regions of non-overlap to ensure restriction to regions where patients had a non-zero probability of receiving either treatment[3] is especially important when considering weighting based on the propensity score. Probabilities close to 0 or 1 could result in large weights that unduly influence the analysis by over-representing patients in atypical circumstances who were certain to receive one of the two treatments. If a large portion of the sample is lost after trimming regions of non-overlap, it could indicate insufficient overlap between distributions. Furthermore, exclusion of observations through trimming because of non-overlap can lead to important changes in the composition of the study population and therefore, could alter the target of inference. In the case example, we assessed propensity score distributional overlap between

doi: 10.1136/bmj.l5657 | *BMJ* 2019;367:l5657 | the bmj

Box 1: Recommended diagnostics and reporting practices for studies using a propensity score weighting method for confounding adjustment

- Evaluate the weight distribution, and consider weight truncation or trimming when extreme weights are encountered
- Describe the study population overall to clearly identify the population for which inference is being made
- Describe the population by exposure groups to evaluate balance achieved across included covariates between treated and reference groups. Consider reporting an overall measure of balance in the weighted sample such as the post weighting C statistic
- Report the crude and weighted effect estimates along with confidence intervals calculated using robust variance that accounts for weighting.

the dabigatran and warfarin groups and noted substantial overlap between the two groups (fig 3). Trimming non-overlapping regions of the propensity score distribution resulted in the exclusion of only 10 patients, which confirmed sufficient overlap.

Evaluating the propensity score distribution in the treatment and reference groups further revealed that the distribution was bimodal for the warfarin group. The first peak comprised of a subset of the warfarin initiators who have a low probability of receiving dabigatran, while the second peak comprised of remaining warfarin initiators who have a relatively higher probability of receiving dabigatran. Examining the distribution after applying weights under different approaches suggested that the patients receiving warfarin in the first peak were down-weighted substantially under all weighting approaches except for the weights targeting the ATE (IPTW and fine stratification weights (ATE)). If the investigators deem that it is important to generate inference that is applicable to all patients with atrial fibrillation initiating dabigatran or warfarin, then it may be appropriate to use weighting approaches that target the ATE in the whole population. However, if investigators consider patients receiving warfarin in the first peak to be a special group of patients with atrial fibrillation where there is little uncertainty over treatment choice (that is, warfarin is always preferred over dabigatran), then it may be appropriate to target the ATT or ATE in the overlap population.

Step 3a: (If sufficient overlap in the propensity score distribution in step 2) Selection of target of inference

As different approaches for weighting based on the propensity score result in estimates targeting different populations, investigators should pay close attention to their target of inference and select a corresponding weighting approach.

Average treatment effect (ATE) in the whole population

Two weighting approaches are available for targeting the ATE, both of which aim to make the distribution of covariates in the treated and reference groups similar to each other and similar to the distribution of the overall study sample.

Inverse probability treatment weighting (IPTW)—This method involves weighting by the inverse probability of receiving the study treatment actually received (1/propensity score for the treated group and 1/(1−propensity score) for the reference group). As the propensity score is directly used to create weights, extreme weights are commonly observed whenever the propensity score is near 0 for a treated patient or near 1 for a reference patient. Weight truncation, which is commonly implemented by setting the maximum and minimum weights at prespecified values based on the observed distribution (eg, 1st and 99th percentile), is routinely necessary to address extreme weights and prevent variance inflation.[16] Although selecting the cutoff value for truncation is often an arbitrary decision, researchers must appreciate that weight truncation involves a bias-variance trade off where truncating more observations by setting a lower threshold (eg, 95th *v* 99th percentile) will further reduce variance inflation, but at a cost of added bias.[28]

In the case example, IPTW as high as 155 417 was observed; truncation at the 99th percentile of the weight distribution led to a maximum weight of 9.91. Another solution to prevent extreme weights is stabilisation by incorporating the marginal probability of receiving the treatment actually received in the numerator.[29] However, stabilising weights in this manner might not completely address all extreme weights, making truncation necessary. In our case example, incorporating marginal probabilities still led to weights of over 100 in 49 observations (>1000 in 20 observations).

A special setting where IPTW is routinely used is in marginal structural modelling.[30] Marginal structural models are particularly useful when accounting for time-varying confounding, formally defined as confounding induced by outcome risk factors that are affected by previous treatment and affect future treatment. In this setting, IPTW calculated at multiple time points throughout the follow-up period are commonly combined with inverse probability of censoring weights to address time-varying confounding and selection bias introduced by informative censoring in a single model.[30] Previously published articles provide additional details on this method.[28 31]

Fine stratification weights targeting the average treatment effect (ATE)—This method does not use the propensity score directly to calculate weights; instead, propensity scores are used to create fine stratums.[9] Stratums can be created in several ways, based on the following:

- The propensity score distribution of the whole cohort
- The propensity score distribution of the smaller of the two exposure groups
- A fixed width of probabilities (eg, 0-0.02 stratum 1, >0.02-0.04 stratum 2, and so on).

For low exposure prevalence, the approach of creating stratums based on the propensity score distribution of the exposed patients ensures assignment of all

Table 2 | Selected patient characteristics before and after propensity score weighting, in case example of dabigatran (D) versus warfarin (W) initiation for atrial fibrillation. Data are number (%) or patients unless stated otherwise

Characteristic	Crude		IPTW* (W; D)	Fine stratified ATE weights (W; D)	Fine stratified ATT weights (W; D)	SMRW (W; D)	Matching weights (W; D)	Overlap weight (W; D)
	W; D	Total						
Weighted (No)	56 456; 22 809	79 255	79 040; 69 264	56 455; 22 800	56 455; 22 800	22 585; 22 800	21 021; 21 256	13 718; 13 717
Age (mean (SD))	71.10 (12.13); 67.29 (12.23)	70.00 (12.28)	69.99 (12.45); 69.84 (11.98)	69.87 (12.42); 70.16 (11.92)	66.81 (12.60); 67.30 (12.22)	67.23 (12.82); 67.30 (12.22)	68.30 (12.26); 68.28 (11.77)	69.06 (12.37); 69.06 (11.90)
Female sex	22 229 (39.4); 8209 (36.0)	30 438 (38.4)	30 464 (38.5); 26 769 (38.6)	21 688 (38.4); 8938 (39.2)	20 350 (36.0); 8205 (36.0)	8235 (36.5); 8205 (36.0)	7794 (37.1); 7837 (36.9)	5179 (37.8); 5178 (37.8)
Coronary artery disease	19 717 (34.9); 6768 (29.7)	26 485 (33.4)	26 504 (33.5); 23 117 (33.4)	18 871 (33.4); 7933 (34.8)	16 776 (29.7); 6766 (29.7)	6787 (30.0); 6766 (29.7)	6447 (30.7); 6464 (30.4)	4317 (31.5); 4317 (31.5)
Systemic embolism	728 (1.3); 112 (0.5)	840 (1.1)	847 (1.1); 640 (0.9)	602 (1.1); 287 (1.3)	289 (0.5); 112 (0.5)	119 (0.5); 112 (0.5)	117 (0.6); 111 (0.5)	88 (0.6); 88 (0.6)
Deep vein thrombosis	4241 (7.5); 289 (1.3)	4530 (5.7)	4533 (5.7); 2014 (2.9)	3260 (5.8); 940 (4.1)	831 (1.5); 289 (1.3)	292 (1.3); 289 (1.3)	291 (1.4); 289 (1.4)	243 (1.8); 243 (1.8)
Pulmonary embolism	2932 (5.2); 103 (0.5)	3035 (3.8)	3035 (3.8); 897 (1.3)	2200 (3.9); 481 (2.1)	388 (0.7); 103 (0.5)	103 (0.5); 103 (0.5)	103 (0.5); 103 (0.5)	94 (0.7); 94 (0.7)
Heart failure	12 464 (22.1); 3648 (16.0)	16 112 (20.3)	16 159 (20.4); 13 893 (20.1)	11 476 (20.3); 4899 (21.5)	9033 (16.0); 3648 (16.0)	3696 (16.4); 3648 (16.0)	3572 (17.0); 3544 (16.7)	2454 (17.9); 2454 (17.9)
Ischaemic stroke	5144 (9.1); 1599 (7.0)	6743 (8.5)	6778 (8.6); 6053 (8.7)	4813 (8.5); 2144 (9.4)	3995 (7.1); 1599 (7.0)	1634 (7.2); 1599 (7.0)	1571 (7.5); 1551 (7.3)	1072 (7.8); 1072 (7.8)
Transient ischaemic attack	2637 (4.7); 947 (4.2)	3584 (4.5)	3586 (4.5); 3139 (4.5)	2556 (4.5); 1070 (4.7)	2356 (4.2); 946 (4.1)	949 (4.2); 946 (4.1)	897 (4.3); 896 (4.2)	596 (4.3); 595 (4.3)
Myocardial infarction	2793 (4.9); 886 (3.9)	3679 (4.6)	3706 (4.7); 3259 (4.7)	2638 (4.7); 1186 (5.2)	2254 (4.0); 885 (3.9)	913 (4.0); 885 (3.9)	861 (4.1); 852 (4.0)	580 (4.2); 580 (4.2)
Peripheral vascular disease or surgery	2675 (4.7); 665 (2.9)	3340 (4.2)	3353 (4.2); 2815 (4.1)	2379 (4.2); 1057 (4.6)	1645 (2.9); 665 (2.9)	678 (3.0); 665 (2.9)	660 (3.1); 652 (3.1)	465 (3.4); 465 (3.4)
Diabetes	14 242 (25.2); 4774 (20.9)	19 016 (24.0)	18 988 (24.0); 16 271 (23.5)	13 526 (24.0); 5594 (24.5)	11 753 (20.8); 4772 (20.9)	4746 (21.0); 4772 (20.9)	4526 (21.5); 4550 (21.4)	3034 (22.1); 3033 (22.1)
Chronic renal disease	6864 (12.2); 1276 (5.6)	8140 (10.3)	8181 (10.4); 6607 (9.5)	5779 (10.2); 2493 (10.9)	3094 (5.5); 1276 (5.6)	1318 (5.8); 1276 (5.6)	1299 (6.2); 1270 (6.0)	980 (7.1); 980 (7.1)

ATE=average treatment effect; ATT=average treatment effect among the treated population; IPTW=inverse probability treatment weights; SMRW=standardised mortality ratio weighting; SD=standard deviation.
*Weights were truncated at the 99th percentile.

exposed individuals to stratums and minimises loss of information. Following stratification, weights for both treated and reference patients in all stratums with at least one treated patient and one reference patient are subsequently calculated based on the total number of patients within each stratum. Stratums with no exposed or reference patients are dropped out before weight calculation. As long as an appropriate stratification procedure is selected to avoid sparse stratums, extreme weights due to propensity scores that are very close to 0 or 1 are unlikely, which is an important strength in circumstances where exposure prevalence is low and propensity score distribution is skewed. These weights are mathematically equivalent to marginal mean weights described in the education literature.[32]

Average treatment effect among the treated population (ATT)

Two weighting approaches are available for targeting the ATT, both of which aim to make the distribution of covariates in the reference group similar to the distribution observed in the treatment group.

Standardised mortality ratio weighting (SMRW)—This method involves setting weights to 1 for the treated patients and weighting reference patients by the odds of treatment probability: (propensity score/(1−propensity score)).[29] Similar to IPTW, SMRW is potentially vulnerable to extreme weights because the propensity score is used directly for calculating the weights. Weight truncation could be considered if large weights are observed.

Fine stratification weights targeting the average treatment effect among the treated population (ATT)—Similar to the fine stratification weights targeting the ATE, propensity scores are used to create fine stratums, but weights for the treated group are set to 1 and reference patients are reweighted based on the number of treated patients residing within their stratum, so that reference patients contribute proportionally to the relative number of total patients within a stratum.[9] Extreme weights are uncommon because propensity score is not directly used to weight but still possible if some stratums are highly imbalanced with respect to the number of treated and reference patients.[33]

Average treatment effect (ATE) in a subset with clinical equipoise

The next two weighting approaches, matching weights and overlap weights, have a variable target of inference that is heavily influenced by overlap in the propensity score distribution. Broadly, these approaches target the ATE in a subset of the overall population with some clinical equipoise. In other words, these approaches aim to make the distribution of covariates in the treated and reference group similar to each other and similar to the distribution in a subset of the overall study sample where patients are eligible to receive either the treatment of interest or the reference treatment.

Matching weights—This method involves weighting patients based on a ratio of the lower of the two predicted probabilities to the predicted probability of

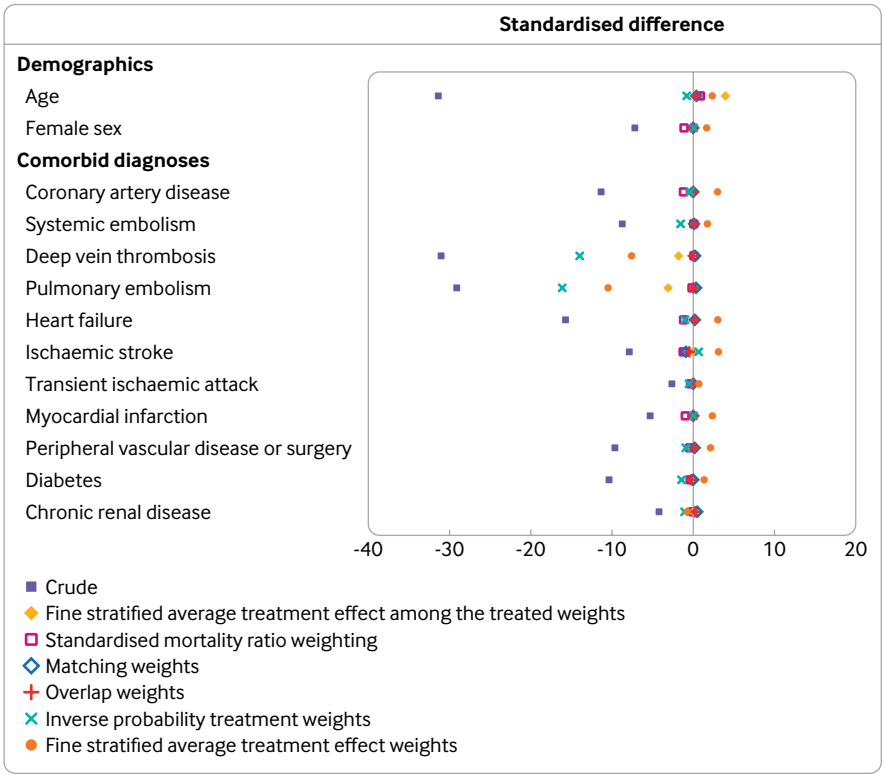

Fig 2 | Standardised differences before and after propensity score weighting, in case example of dabigatran versus warfarin initiation for atrial fibrillation, by selected patient characteristics

the actually received treatment.[10 11] A key feature is that extreme weights are impossible because weights are bound between 0 and 1 by design, eliminating the need for weight truncation. The target of inference is close to the ATE in the whole population when groups are equally sized, and propensity score distributions have good overlap and is close to the ATT in the group with fewer observations when groups are unequally sized, but propensity score distributions have good overlap. In circumstances of limited overlap in propensity score distribution, this approach targets treatment effect estimation in a subpopulation that is neither the set of patients receiving the treatment of interest in routine care nor the whole study population.

Overlap weights—This method involves weighting patients based on the predicted probability of receiving the opposite treatment.[12] Similar to matching weights, extreme weights are impossible as weights are bound between 0 and 1 by design and, therefore, no truncation is necessary. Further, an attractive feature is that this weighting method yields exact covariate balance between treated and reference groups by construction. However, the target of inference is the ATE in the overlap population, which might be different from the ATT or the ATE in the whole study population.

For the case example, we calculated the treatment effect comparing dabigatran and warfarin for the risk of major bleeding before and after weighting for all approaches. The results are reported in figure 4, along with confidence intervals calculated using robust variance estimators. The crude estimate suggested a substantially lower bleeding risk with dabigatran versus warfarin, which attenuated after adjustment for confounding through all weighting approaches. Overall, hazard ratio estimates for approaches with a similar target of inference were nearly identical. Hazard ratios for approaches targeting the ATE and ATT were somewhat different (0.73 *v* 0.79). One potential explanation of this difference could be effect measure modification by patient characteristics. Because these estimates apply to populations with varying distribution of patient characteristics (as seen in table 1), presence of effect measure modification could lead the estimates to diverge.

Step 3b: (If insufficient overlap in the propensity score distribution in step 2) Consider alternative comparison groups or other design modifications

Insufficient distributional overlap could indicate two treatments that are used in completely different populations or for different indications. In this circumstance, investigators should reconsider their design choices with respect to the comparison group or study inclusion criteria. If sufficient overlap is achieved after such modifications, then use of weighting based on the propensity score could be considered, based on the considerations summarised in step 3a. If alternative comparison groups or design modifications fail to achieve sufficient overlap, investigators might need to reconsider the study question.

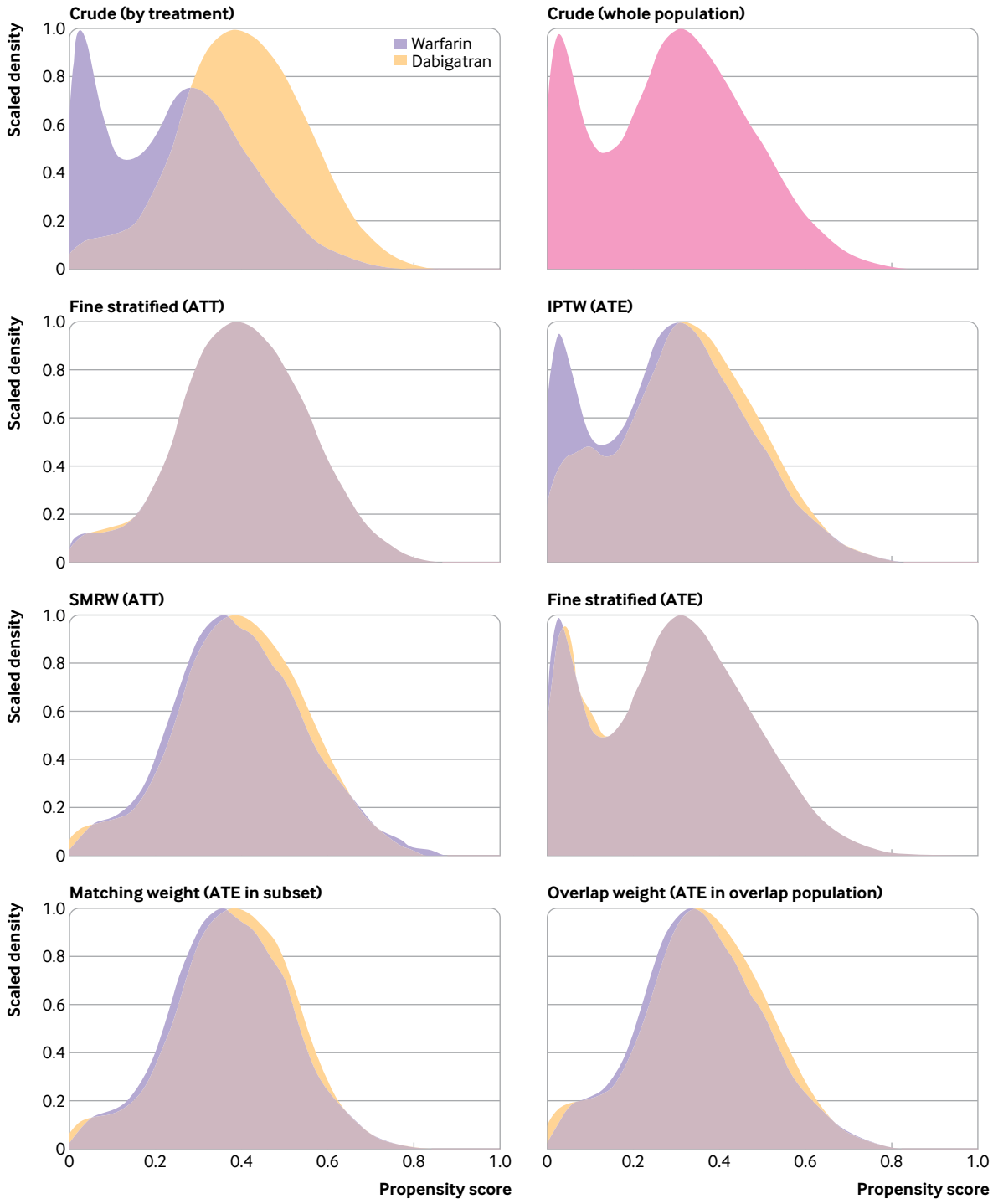

Fig 3 | Propensity score distributional overlap before and after propensity score weighting, in case example of dabigatran versus warfarin initiation for atrial fibrillation. ATE=average treatment effect; ATT=average treatment effect among the treated population; IPTW=inverse probability treatment weights; SMRW=standardised mortality ratio weighting

Propensity score based weighting approaches for confounding adjustment in evaluations of comparative outcomes in more than two treatment groups

Certain weighting approaches readily extend to settings of more than two treatment groups. Specifically, weight calculations for IPTW, matching weights, and SMRW in settings of two groups have direct equivalents for settings of three or more treatment groups. All these approaches involve generating propensity scores for three or more treatments in a multinomial logistic regression model. IPTWs are calculated based on the inverse of the propensity of the treatment actually received, and target ATE in the whole population regardless of the number of treatment groups. For matching weights in settings of three or more groups, the numerator includes the minimum of all

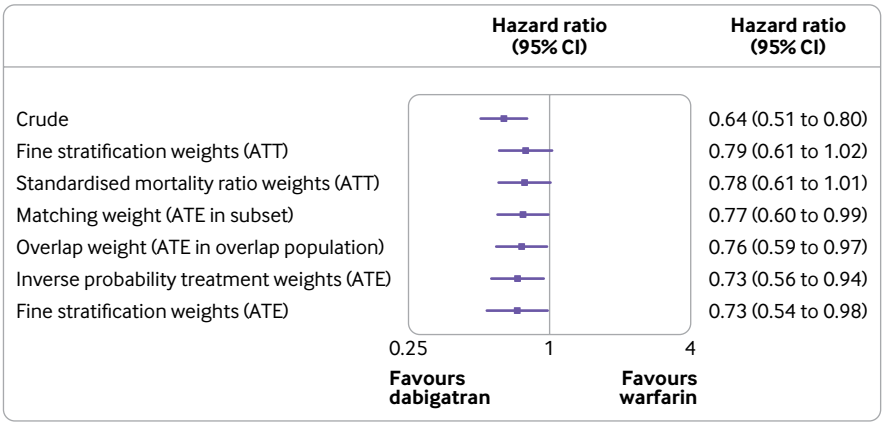

Fig 4 | Hazard ratios (95% confidence intervals) for dabigatran versus warfarin for the risk of ischaemic stroke or systemic embolism, by approach of propensity score based weighting. ATE=average treatment effect; ATT=average treatment effect among the treated population

available propensity scores for each patient and the denominator includes propensity of the treatment actually received.[11] Similar to settings of two treatment groups, when treatment groups are equally sized and covariate overlap is substantial across three or more treatment groups, matching weights target ATE in the whole population; when one of the treatment groups is small and covariate overlap is substantial, matching weights target ATT in the smallest group.[11] For SMRW, investigators can target ATT for a specific treatment group by setting weights for patients receiving the target treatment to 1 and calculating weights for other treatment groups as a ratio of propensity of the target treatment to propensity of the treatment actually received. An extension of overlap weights, termed as generalised overlap weights, has been proposed for settings of three or more groups where weights are constructed as the product of the inverse probability weights and the harmonic mean of the generalised propensity scores and these weights target the population with the most overlap in covariates across the multiple treatments.[13] Extension to settings of three or more groups for the weighting approaches based on fine stratification requires simultaneous stratification on a multinomial propensity score, which would increase the number of stratums exponentially and could result in variable estimates.[34]

Conclusion

Weighting based on the propensity score represents a valuable tool for confounding adjustment in observational studies of treatment use and is increasingly being used in epidemiological investigations. In this article, we outline key considerations involved in selection and implementation of an appropriate weighting approach based on the propensity score that could provide a framework for practitioners in designing and reporting their analysis.

Contributors: RJD and JMF have jointly developed this manuscript. RJD is the guarantor of the content of this article.

Funding: This study was supported through internal funding from the Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Harvard Medical School/Brigham and Women's Hospital.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: support from Harvard Medical School/Brigham and Women's Hospital for the submitted work; RJD is principal investigator of research grants from Bayer, Novartis, and Vertex, to the Brigham and Women's Hospital for unrelated work; no other relationships or activities that could appear to have influenced the submitted work.

Data sharing: Patient level data are not made available publicly according the data use agreement. Any aggregate level data not presented in the manuscript can be requested from the corresponding author.

The lead author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

1 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41-55. doi:10.1093/biomet/70.1.41
2 Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc* 2005;100:322-31. doi:10.1198/016214504000001880
3 Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol* 2006;98:253-9. doi:10.1111/j.1742-7843.2006.pto_293.x
4 Stürmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol* 2005;161:891-8. doi:10.1093/aje/kwi106
5 Kurth T, Walker AM, Glynn RJ, et al. 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. doi:10.1093/aje/kwj047
6 Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. 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. *J Clin Epidemiol* 2006;59:437-47. doi:10.1016/j.jclinepi.2005.07.004
7 King G, Nielsen R. Why propensity scores should not be used for matching. https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching
8 Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. *Stat Med* 2014;33:4053-72. doi:10.1002/sim.6207
9 Desai RJ, Rothman KJ, Bateman BT, Hernandez-Diaz S, Huybrechts KF. A propensity-score-based fine stratification approach for confounding adjustment when exposure is infrequent. *Epidemiology* 2017;28:249-57. doi:10.1097/EDE.0000000000000595
10 Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat* 2013;9:215-34. doi:10.1515/ijb-2012-0030

11 Yoshida K, Hernández-Díaz S, Solomon DH, et al. Matching weights to simultaneously compare three treatment groups: comparison to three-way matching. *Epidemiology* 2017;28:387-95. doi:10.1097/EDE.0000000000000627
12 Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol* 2019;188:250-7.
13 Li F. Propensity score weighting for causal inference with multi-valued treatments. arXiv 180805339 [Preprint]. 2018. https://arxiv.org/abs/1808.05339
14 Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med* 2013;32:2837-49. doi:10.1002/sim.5705
15 Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc* 1987;82:387-94. doi:10.1080/01621459.1987.10478441
16 Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* 2015;34:3661-79. doi:10.1002/sim.6607
17 Lauffenburger JC, Farley JF, Gehi AK, Rhoney DH, Brookhart MA, Fang G. Effectiveness and safety of dabigatran and warfarin in real-world US patients with non-valvular atrial fibrillation: a retrospective cohort study. *J Am Heart Assoc* 2015;4:e001798. doi:10.1161/JAHA.115.001798
18 Huybrechts KF, Hernández-Díaz S, Patorno E, et al. Antipsychotic use in pregnancy and the risk for congenital malformations. *JAMA Psychiatry* 2016;73:938-46. doi:10.1001/jamapsychiatry.2016.1520
19 Desai RJ, Wyss R, Jin Y, et al. Extension of disease risk score-based confounding adjustments for multiple outcomes of interest: an empirical evaluation. *Am J Epidemiol* 2018;187:2439-48. doi:10.1093/aje/kwy130
20 Wyss R, Ellis AR, Brookhart MA, et al. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *Am J Epidemiol* 2014;180:645-55. doi:10.1093/aje/kwu181
21 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. doi:10.1002/pds.1555
22 Wyss R, Girman CJ, LoCasale RJ, Brookhart AM, Stürmer T. Variable selection for propensity score models when estimating treatment effects on multiple outcomes: a simulation study. *Pharmacoepidemiol Drug Saf* 2013;22:77-85. doi:10.1002/pds.3356
23 Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol* 2011;174:1213-22. doi:10.1093/aje/kwr364
24 Waernbaum I. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Stat Med* 2012;31:1572-81. doi:10.1002/sim.4496
25 Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One* 2011;6:e18174. doi:10.1371/journal.pone.0018174
26 Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat Simul Comput* 2009;38:1228-34. doi:10.1080/03610910902859574
27 Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. *Stat Med* 2014;33:1685-99. doi:10.1002/sim.6058
28 Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008;168:656-64. doi:10.1093/aje/kwn164
29 Brookhart MA, Wyss R, Layton JB, Stürmer T. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes* 2013;6:604-11. doi:10.1161/CIRCOUTCOMES.113.000359
30 Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550-60. doi:10.1097/00001648-200009000-00011
31 Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. *Int J Epidemiol* 2017;46:756-62.
32 Hong G. Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *J Educ Behav Stat* 2010;35:499-531. doi:10.3102/1076998609359785
33 Franklin JM, Eddings W, Austin PC, Stuart EA, Schneeweiss S. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Stat Med* 2017;36:1946-63.
34 Desai RJ, Yoshida K, Huybrechts K, Franklin JM. An empirical evaluation of a propensity score stratification based weighting approach in settings of more than two treatment groups. *Pharmacoepidemiol Drug Saf* 2019;28:19-19.