

WHAT IS PS?

- Propensity score based methods target causal inference in observational studies to mimic randomised experiments by facilitating the measurement of differences in outcomes between the treated population and a reference population
- 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.

CONFOUNDING ADJUSTMENT BY PS

- 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.
- Advantages
 - 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

PS MATCH

- Common used
- Limitation: discarding unmatched observations falling within the caliper after a prespecified number of observations
- increasing covariate imbalance after propensity score matching has been described by King and Nielsen
- Other PS: stratification, adjustment as a regressor, and weighting are not affected by this paradox

WEIGHTING

- weighting keeps most observations in the analysis offer increased precision when estimating treatment effects.
- Easily to transparent reporting of the balance achieved between treatment and reference populations
- Most flexible approach in the analysis with multiple available variations that allow targeting specific populations for inference
- Tradition: inverse probability treatment weights (IPTW) or standardised mortality ratio weights (SMRW)
- newer approaches (including propensity score fine stratification weights, matching weights, and overlap weights)

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 / stratification achieve balance by ensuring that treated and reference populations on average have comparable propensity scores
- 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*

BASIC PRINCIPLE OF WEIGHTING METHODS BASED ON PROPENSITY SCORES (2)

- 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
- Sandwich type estimator is recommended for variance estimation for the treatment effect estimates.

TARGET OF INFERENCE (ESTIMAND)

- Would it be feasible to treat all eligible patients included in the study with the treatment of interest?
- If Yes, target of inference might be defined as the average treatment effect (ATE).
- Ex. Dabigatran vs warfarin for prevent stroke. Both of these treatments are indicated as exchangeable options
- If No, target of inference might be defined as average treatment effect among the treated population (ATT).
- ATT: only patients with certain characteristics who actually received the treatment would be ideal candidates for treatment
- Ex. Antipsychotic in pregnancy, only women with greater severity of these conditions would receive treatment with antipsychotics during pregnancy
- In the absence of treatment effect heterogeneity by patient characteristics, ATE and ATT will coincide.

WHEN SELECTING THE WEIGHTING METHOD

- Step 1: Correct specification of the propensity score model
- Avoiding misspecification of the propensity score model
- 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
- The covariate balancing propensity scores or machine learning approaches such as neural networks— could provide alternatives that are less prone to misspecification

STEP 1: CORRECT SPECIFICATION OF THE PS MODEL

- Regardless of the approach, researchers should emphasise inclusion of outcome risk factors in the model and exclusion of strong predictors of treatment that are not associated with outcomes (instrumental variable) to avoid increased variance and amplification of bias due to unmeasured confounding
- Approaches that use the score directly to create weights (IPTW) are theoretically more prone to increased bias and variance from misspecification.
- Stratification might be more robust against misspecification, because it can be conceptualised as a semiparametric implementation of propensity score weighting that uses the score only to create strata and then uses the counts of observations within each stratum to derive weights

STEP 1: CORRECT SPECIFICATION OF THE PS MODEL

- For reporting the balance, a measure such as the standardised difference in prevalence (or means for continuous variables) is recommended, and also considered overall balanced e.g. post weighting C statistics, where values closer to 0.5 indicates overall balance achievement.
- Ex. Y = initiation of dabigatran
X = 66 pt characteristics
logistic regression model
Condition on Ps via weight approaches led to balance among covariates

STEP 2: EVALUATION OF PS DISTRIBUTIONAL OVERLAP

- High overlap in the propensity score distribution generally indicates a reasonable degree of clinical equipoise in treatment selection.
- The recommendation is *trimming the regions of non-overlap* to ensure restriction to regions where patients had a nonzero probability of receiving either Tx is important when considering weighting based on the PS.
- Probabilities close to 0 or 1 could result in large weights that unduly influence the analysis by over-representing patients who were certain to receive one of the two treatments
- If a large portion of the sample is lost after trimming regions of nonoverlap, it could indicate insufficient overlap between distributions.
- 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.

- 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))

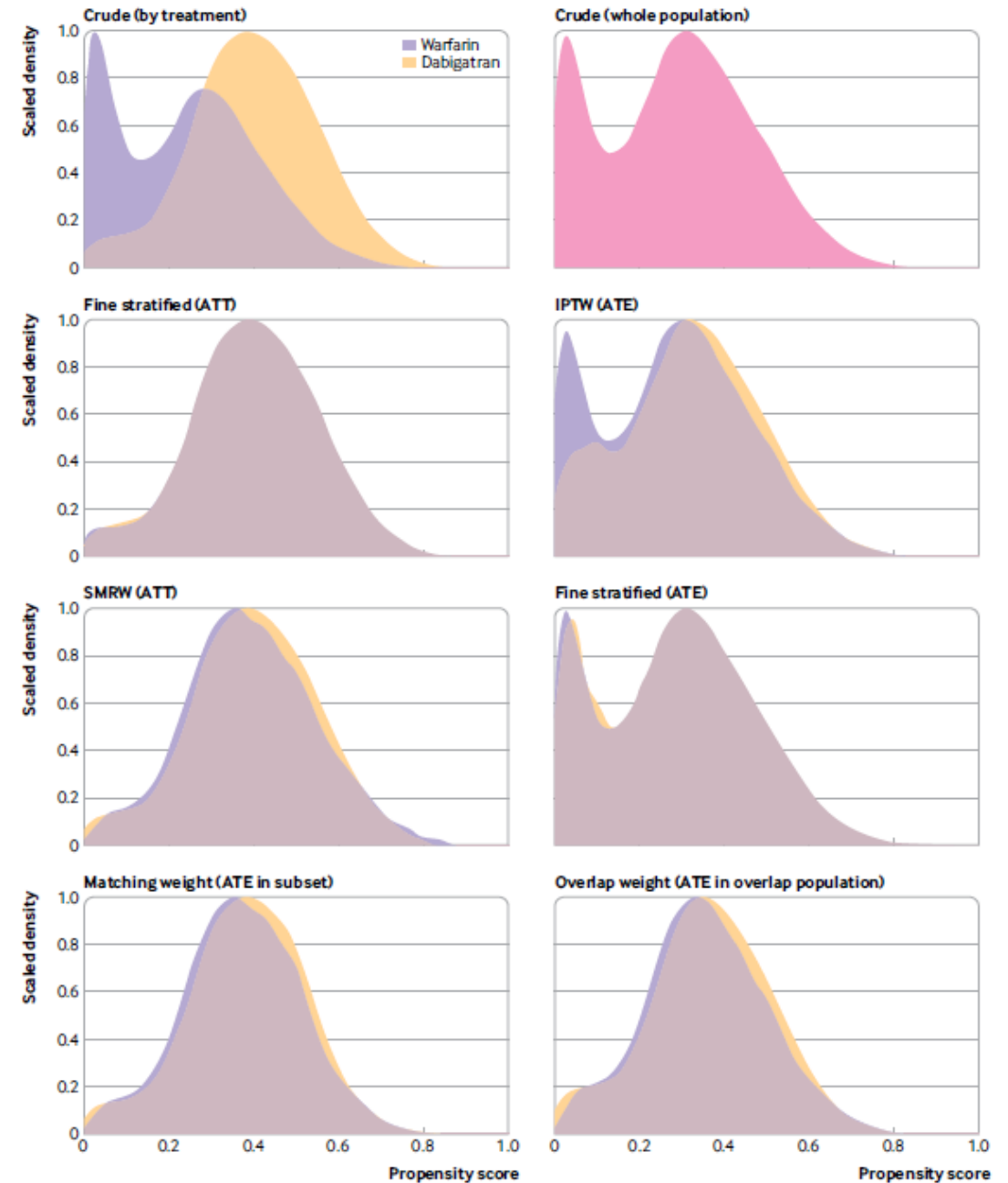


Fig 3 | Propensity score distributional overlap before and after propensity score weighting. In case example of dabigatran versus warfarin initiation for atrial fibrillation. ATE=average treatment effect; ATT=average treatment effect among the treated population; IPTW=inverse probability treatment weights;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

STEP 3A: (IF SUFFICIENT OVERLAP) SELECTION OF TARGET OF INFERENCE

- As different approaches for weighting based on the PS result in estimates targeting different populations
- ATE in whole population
- 2 weight approaches for ATE, 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.
 - (IPTW and fine
 - stratification weights (ATE)).

ATE: IPTW

- Weighting by the inverse probability of receiving the study treatment actually received ($1/\text{propensity score}$ for the treated group and $1/(1-\text{propensity score})$ for the reference group).
- Propensity score is directly used to create weights, extreme weights are commonly observed whenever the PS 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

ATE: IPTW: EXTREMELY WEIGHT

- 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.
- Another solution to prevent extreme weights is stabilisation by incorporating the marginal probability of receiving the treatment actually received in the numerator. ???, however, might not completely address all extreme weights, making truncation necessary.

ATE: IPTW: MARGINAL STRUCTURAL MODEL

- 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.
- 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

ATE: FINE STRATIFICATION WEIGHTS TARGETING THE AVERAGE TREATMENT EFFECT (ATE)

- PS are used to create fine strata instead of directly calculate weights, based on 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 stratum1, >0.02-0.04 stratum 2, and so on).
- For low exposure prevalence, the approach of creating strata based on the propensity score distribution of the exposed patients ensures assignment of all exposed individuals to strata and minimises loss of information.

ATE: FINE STRATIFICATION WEIGHTS

- Following stratification, weights for both treated and reference patients in all strata with at least one treated patient and one reference patient are subsequently calculated based on the total number of patients within each stratum.
- Strata with no exposed or reference patients are dropped out before weight calculation
- An *appropriate stratification* procedure is selected to *avoid sparse strata*, 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*

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
 - Fine stratification weights targeting ATT

ATT: 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)).
- 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 ATT

- Propensity scores are used to create fine strata, 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.
- Then reference patients contribute proportionally to the relative number of total patients within a stratum
- Extreme weights are uncommon because propensity score is not directly used to weight but still possible if some strata are highly imbalanced with respect to the number of treated and reference patients.

ATE IN A SUBSET WITH CLINICAL EQUIPOISE

- Matching weights and overlap weights approach target the ATE in a subset of the overall population with some clinical equipoise
- have a variable target of inference that is heavily influenced by overlap in the propensity score distribution
- These approach 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

ATE IN CLINICAL EQUIPOISE: MATCHING WEIGHTS

- This method involves weighting patients based on a *ratio of the lower of the two predicted probabilities ??* to the predicted probability of the actually received treatment.
- A key feature is that extreme weights are impossible because weights are bound between 0 and 1 by design (no need 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
 - is close to the ATT in the group with fewer observations when groups are unequally sized, but propensity score distributions have good overlap.

ATE IN CLINICAL EQUIPOISE: OVERLAP WEIGHTS

- This method involves weighting patients based on the predicted probability of receiving the opposite treatment
- Extreme weights are impossible as weights are bound between 0 and 1 by design (no need truncation)
- 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.

CRUDE VS PS BASED WEIGHTING

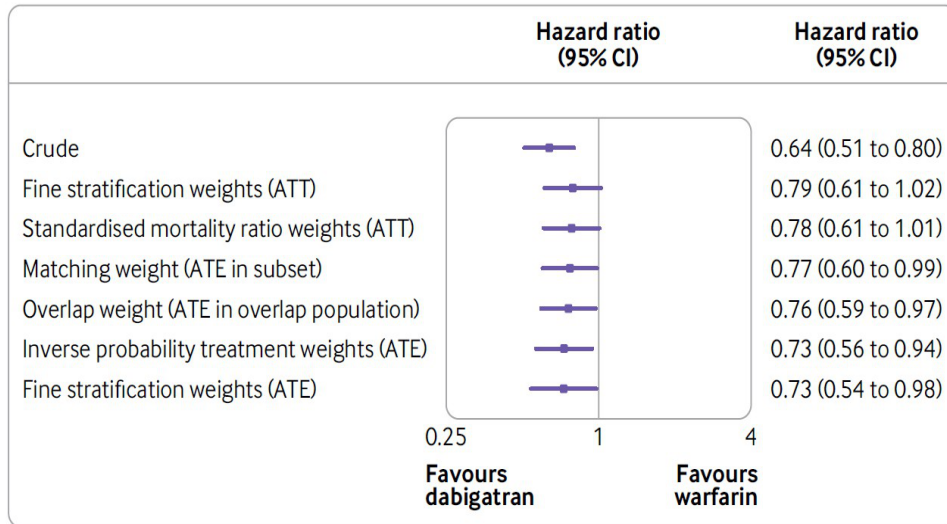


Fig 4 | Hazard ratios (95% confidence intervals) for dabigatran versus warfarin for the risk of ischaemic stroke or systemic embolism, by approach of propensity score based weighting. ATE=average treatment effect; ATT=average treatment effect among the treated population

- The crude estimate suggested a substantially lower bleeding risk with dabigatran versus warfarin, which attenuated after adjustment for confounding through all weighting approaches.
- Overall HR 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.



**STEP 3B: (INSUFFICIENT OVERLAP PS
DISTRIBUTION)
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.
- Investigators should reconsider their design choices with respect to the comparison group or study inclusion criteria.
- If alternative comparison groups or design modifications fail to achieve sufficient overlap, investigators might need to reconsider the study question.

PS WEIGHTING IN MORE THAN 2 TX GROUPS

- Weight calculations for IPTW, matching weights, and SMRW in settings of two groups have direct equivalents for settings of three or more treatment groups.
- 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 available propensity scores for each patient and the denominator includes propensity of the treatment actually received

PS WEIGHTING IN MORE THAN 2 TX GROUPS

- 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

Table 1 | Alternative approaches for weighting based on propensity scores

Method	Weight calculation		Target of inference (estimand)	Features	Interpretation
	Treated patients	Reference patients			
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)	$\frac{(N_{\text{total in PS stratum } i} / N_{\text{total}})}{(N_{\text{exposed in PS stratum } i} / N_{\text{total exposed}})}$	$\frac{(N_{\text{total in PS stratum } i} / N_{\text{total}})}{(N_{\text{reference in PS stratum } i} / N_{\text{total reference}})}$	ATE in the whole population	Does not use the PS directly to calculate weights; instead, <u>the scores are used to create fine strata and weights are subsequently calculated</u> 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	

Table 1 | Alternative approaches for weighting based on propensity scores

Method	Weight calculation		Target of inference (estimand)	Features	Interpretation
	Treated patients	Reference patients			
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	$\frac{(N_{\text{exposed in PS stratum } i} / N_{\text{total exposed}})}{(N_{\text{reference in PS stratum } i} / N_{\text{total reference}})}$	ATT	Does not use the PS directly to calculate weights; instead, the scores are used to create fine strata 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	

Table 1 | Alternative approaches for weighting based on propensity scores

Method	Weight calculation		Target of inference (estimand)	Features	Interpretation
	Treated patients	Reference patients			
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

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

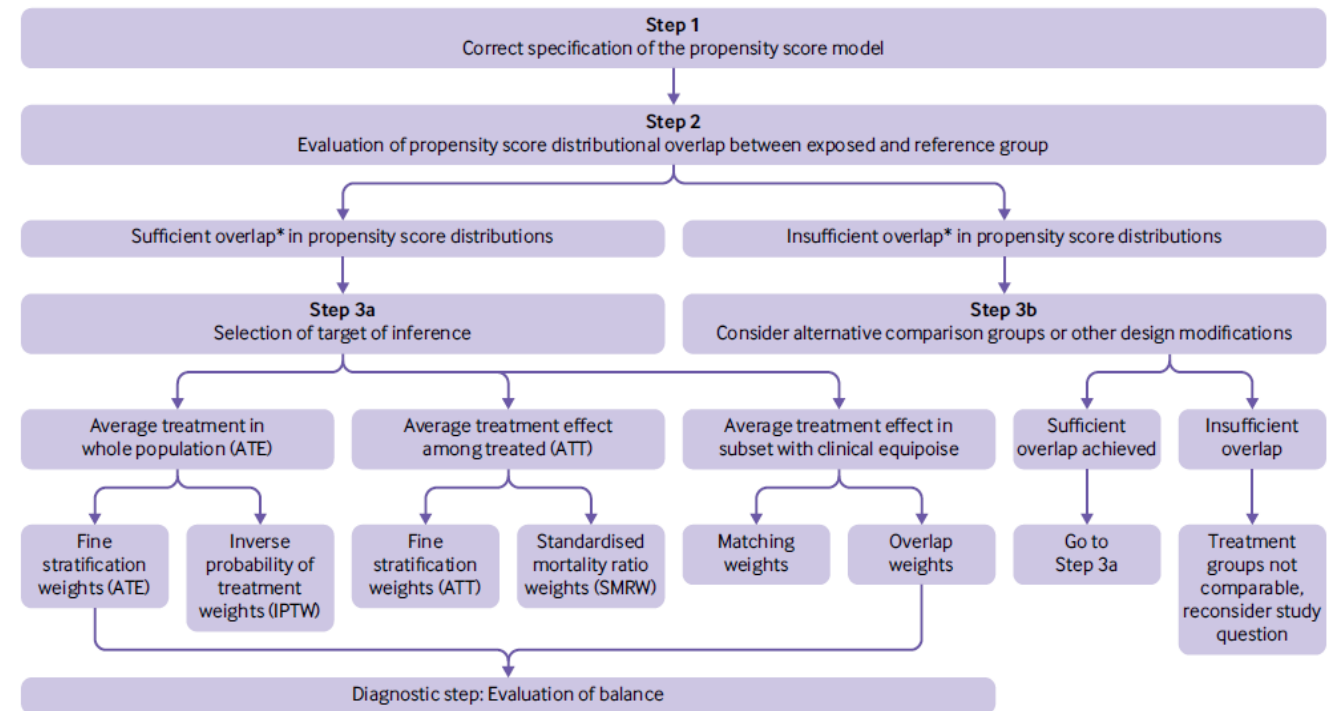
- Evaluate the weight distribution, and consider weight truncation or trimming when extreme weights are encountered
- Describe the study population overall to clearly identify the population for which inference is being made
- Describe the population by exposure groups to evaluate balance achieved across 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.

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

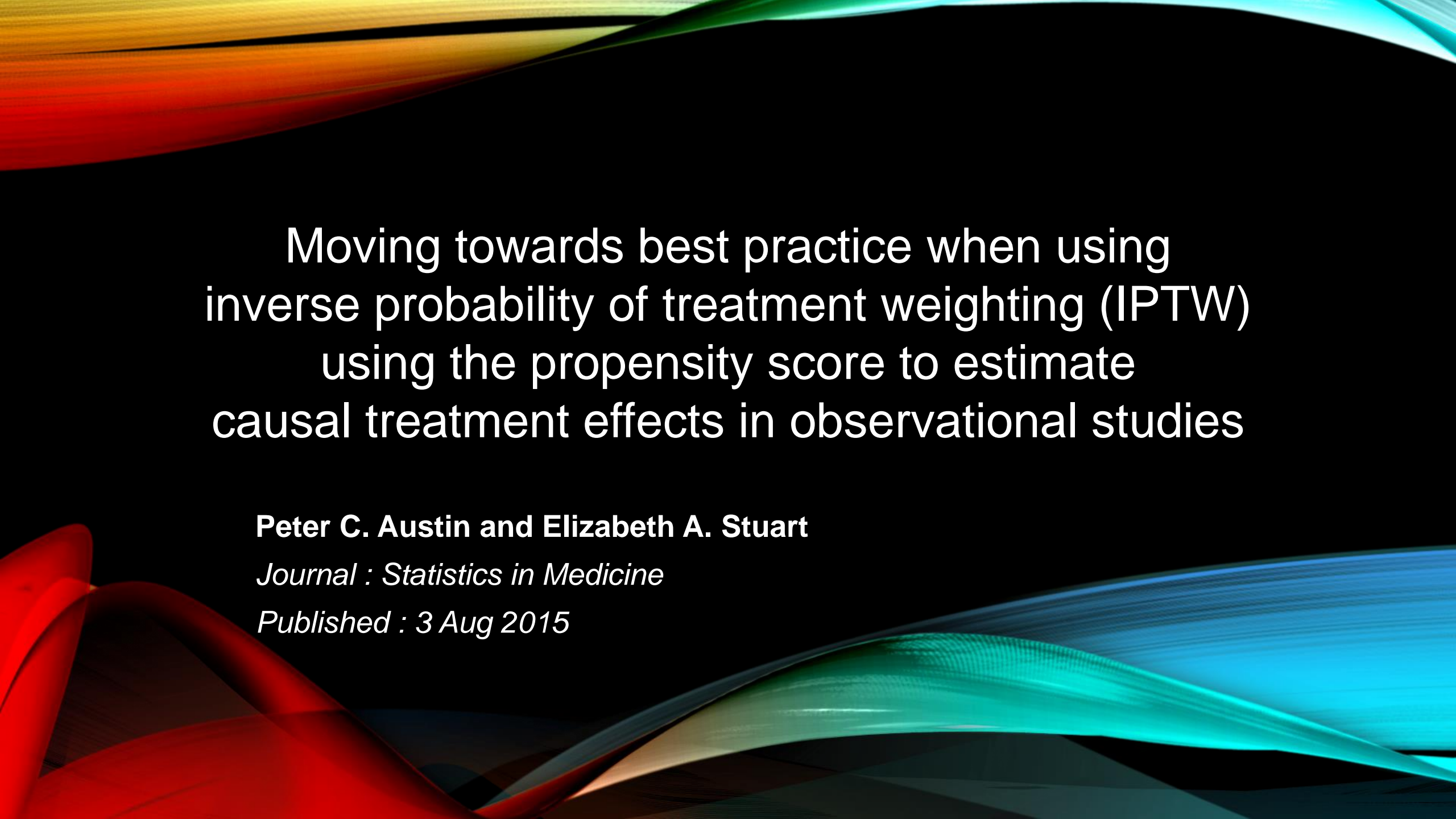
Principles outlined in this report are intended to help investigators in identifying the most suitable propensity score based weighting approach for their analysis and provide a framework for transparent reporting





CRITICIZE

Tunlanut Sapankaew
Commentator



Moving towards best practice when using
inverse probability of treatment weighting (IPTW)
using the propensity score to estimate
causal treatment effects in observational studies

Peter C. Austin and Elizabeth A. Stuart

Journal : Statistics in Medicine

Published : 3 Aug 2015

COMMONLY USED PROPENSITY SCORE METHODS

- Covariate adjustment using the propensity score
- stratification or subclassification on the propensity score
- Matching on the propensity score
- Inverse probability of treatment weighting (IPTW)

INVERSE PROBABILITY OF TREATMENT WEIGHTING

1. *Potential outcomes framework and average treatment effects*
2. *The propensity score and inverse probability of treatment weighting*
3. *Variable selection for the propensity score model*
4. *Assumptions of propensity score methods*

IPTW: POF AND ATE

- The potential outcomes framework assumes that each subject has a pair of potential outcomes: $Y_i(0)$ and $Y_i(1)$, the outcomes under the control treatment and the active treatment when received under identical circumstances
- However, each subject receives only one of the control treatment or the active treatment, thus only one outcome is observed for each subject
- For each subject, the effect of treatment is defined as $Y_i(1) - Y_i(0)$: the difference between the two potential outcomes. The *average treatment effect* (ATE) is defined to be: $E[Y_i(1) - Y_i(0)]$
- The ATE is the average effect, at the population level, of moving an entire population from control to treated.

IPTW: TREATMENT ASSIGNMENT

- If treatment were assigned at random, treatment assignment is independent of the potential outcomes. Thus, randomization provides an unbiased estimate of the average treatment effect.
- However, an observational study simply comparing outcomes between the two treatment groups does not necessarily yield an unbiased estimate of the average treatment effect.

PS AND IPTW

- The inverse probability of treatment weight is defined as each subject's weight is equal to the inverse of the probability of receiving the treatment that the subject received.
- Methods for estimating treatment effects that use weighting by the inverse of the probability of treatment
- ATE can be estimated by
n; number of subjects
- Weighting by IPTW results in an artificial population in which baseline covariates are independent of treatment status
- allows the comparison of distributions between treated and control subjects, estimating variances and significance levels

$$\frac{1}{n} \sum_{i=1}^n \frac{Z_i Y_i}{e_i} - \frac{1}{n} \sum_{i=1}^n \frac{(1-Z_i) Y_i}{1-e_i}$$

CONVENTIONAL WEIGHT VS STABILIZE WEIGHT

- Conventional IPTW weight

$$w = \frac{Z}{e} + \frac{1-Z}{1-e}$$

- Where e = ps score

$$e = P(Z = 1 | \mathbf{X})$$

- Z = treatment assignment
- \mathbf{X} = observed baseline covariates

- Stabilize weight

$$w = \frac{Z \Pr(Z=1)}{e} + \frac{(1-Z) \Pr(Z=0)}{1-e}$$

- $\Pr(Z = 1)$ and $\Pr(Z = 0)$ denote the marginal probability of treatment and control in the overall sample

IPTW: DIFFICULTY AND SOLUTION

- Treated subjects with a very low PS can result in a very large weight. Similarly, a control subject with a PS close to 1 can result in a very large weight, such weights can increase the variability of the estimated treatment effect
- Alternative to address the problems that can arise with very large weights is to use trimmed or truncated weights, in which weights that exceed a specified threshold are each set to that threshold.
- The threshold is often based on quantiles of the distribution of the weights (e.g., the 1st and 99th percentiles).

VARIABLE SELECTION FOR THE PS

- A suggestion is to include variables that influence the treatment selection process, however
- Prior evidence has suggested that, it is preferable to include either the prognostically important covariates (those related to outcomes) or the confounding covariates (those related to treatment and outcomes) in the propensity score model than to include those variables that affect the treatment-selection process.
- By using causal diagrams in conjunction with a review of the subject-matter literature and expert opinion.
- Instruments (i.e., variables that affect treatment-selection but not the outcome) can result in increased bias and variance of the treatment-effect estimate.

IPTW: ASSUMPTIONS

- Causal inference using the propensity score requires four assumptions
 1. *Consistency*: a subject's potential outcome under the treatment actually received is equal to the subject's observed outcome
 2. *Exchangeability* (ignorable treatment assignment): no unmeasured confounders: that one has measured and has access to all of the variables that affect treatment selection and outcomes. (cannot be formally tested)
 3. *Positivity*: all subjects have a non-zero probability of receiving each treatment
 4. *no misspecification*: model should be correctly specified, however, authors suggest focusing on assessing balance of measured covariates between treated and control subjects in the weighted sample.

BALANCE ASSESSMENT FROM 29 ARTICLES

- 3 articles presented some assessment of the distribution of weights
 - One computed mean (SD) of the weights, reported the range of the weights conducted separate analyses using *stabilized weights and trimmed weights*.
 - Second study conducted 3 separate analyses using *conventional weights, standardized weights, and trimmed weights* and used boxplots to examine the distribution of the weights.
- 14 articles assessed the distribution of baseline covariates after implementing IPTW.
 - These included using standardized differences in the weighted sample, the Kolmogorov–Smirnov statistic, a crude comparison of baseline characteristics, and statistical significance testing in the weighted sample.
- 2 articles conducted an assessment of baseline covariate balance and examined the distribution of the weights
- Several sets of authors incorrectly defined the weights as the reciprocal of the propensity score, rather the reciprocal of the probability of receiving the treatment that was actually received.

IPTW DIAGNOSTICS OF BALANCE

1. *Comparison of means and proportions of baseline variables*
2. *Comparison of interactions and higher-order moments of continuous variables*
3. *Graphical comparisons of the distribution of continuous variables*
4. *Numerical comparisons of the distribution of continuous variables.*

IPTW : DIAGNOSTIC OF BALANCE: COMPARISON OF MEANS AND PROPORTIONS OF BASELINE VARIABLES (1)

- In an unweighted sample, the standardized difference for continuous variable is defined as

$$d = 100 \times \frac{(\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}})}{\sqrt{\frac{s_{\text{treatment}}^2 + s_{\text{control}}^2}{2}}}$$

- For dichotomous variables the standardized difference is defined as

$$d = 100 \times \frac{(\hat{p}_{\text{treatment}} - \hat{p}_{\text{control}})}{\sqrt{\frac{\hat{p}_{\text{treatment}}(1-\hat{p}_{\text{treatment}}) + \hat{p}_{\text{control}}(1-\hat{p}_{\text{control}})}{2}}}$$

- The standardized difference compares the difference in means in units of the pooled standard deviation
- The use of the standardized difference (is not influenced by sample size) can be used to compare balance in measured variables between treated and control subjects in the same sample when different weights are assigned to the same subjects.

IPTW : diagnostic of balance: *Comparison of means and proportions of baseline variables (2)*

- While there is no consensus as to what value of a standardized difference can be taken to indicate the presence of meaningful confounding
- Some authors have suggested that a standardized difference in excess of 10% may be indicative of meaningful imbalance in a covariates between treated and control subjects

IPTW : DIAGNOSTIC OF BALANCE: COMPARISON OF INTERACTIONS AND HIGHER-ORDER MOMENTS OF CONTINUOUS VARIABLES.

- Author suggest that standardized differences be used to compare the mean of higher-order moments (e.g., squares and cubes of continuous variables) and interactions between continuous variables
- Comparing the mean of squares of continuous variables is equivalent to comparing the variance of that variable between treatment groups.
- One wants to ensure that the variance, and not only the mean, of a continuous variable is similar between treatment groups in the weighted sample

IPTW : DIAGNOSTIC OF BALANCE: *GRAPHICAL COMPARISONS OF THE DISTRIBUTION OF CONTINUOUS VARIABLES*

- One wants to induce balance on the entire distribution of continuous covariates, not just means and higher-order terms of baseline variables
- Graphical methods i.e. Side-by-side boxplots and empirical cumulative distribution functions (CDFs) can be used to compare the distribution of continuous baseline covariates between treated and control subjects in the weighted sample.
- Side-by-side boxplots to compare the distribution of baseline covariates between treated and control subjects in the weighted sample
- A limitation of the graphical approach is that it relies on a subjective comparison of graphs, especially when comparing two different specifications of the propensity score model

IPTW : DIAGNOSTIC OF BALANCE: *NUMERICAL COMPARISONS OF THE DISTRIBUTION OF CONTINUOUS VARIABLES*

- Authors suggest the Kolmogorov–Smirnov test permit a formal comparison of the distribution of a continuous variable between two independent groups, that permits a quantification of the difference in the distribution of a continuous baseline covariate between treated and control subjects.
- The test statistic is defined to be the maximal vertical distance between the two empirical CDFs of the variable in the two groups

IPTW: POSITIVITY ASSUMPTION TESTING

- Cole and Hernan suggest that analysts should determine the mean stabilized weight
- If the mean of the stabilized weights is far from 1 or if there are very extreme values, then this can be indicative of non-positivity or that the propensity score model has been misspecified.
- The standard deviation of the weights can be useful when comparing between different specifications of the propensity score model.



EXAMPLE

- P: Pt with MI 1999-2001, Canada
 - I: Prescribe beta-blocker at D/C
 - C: no beta-blocker at D/C
 - O: CVD
- Eighteen of the 24 measured baseline covariates had standardized differences that exceeded 10% may be indicative of meaningful imbalance in a covariates between treated and control subjects

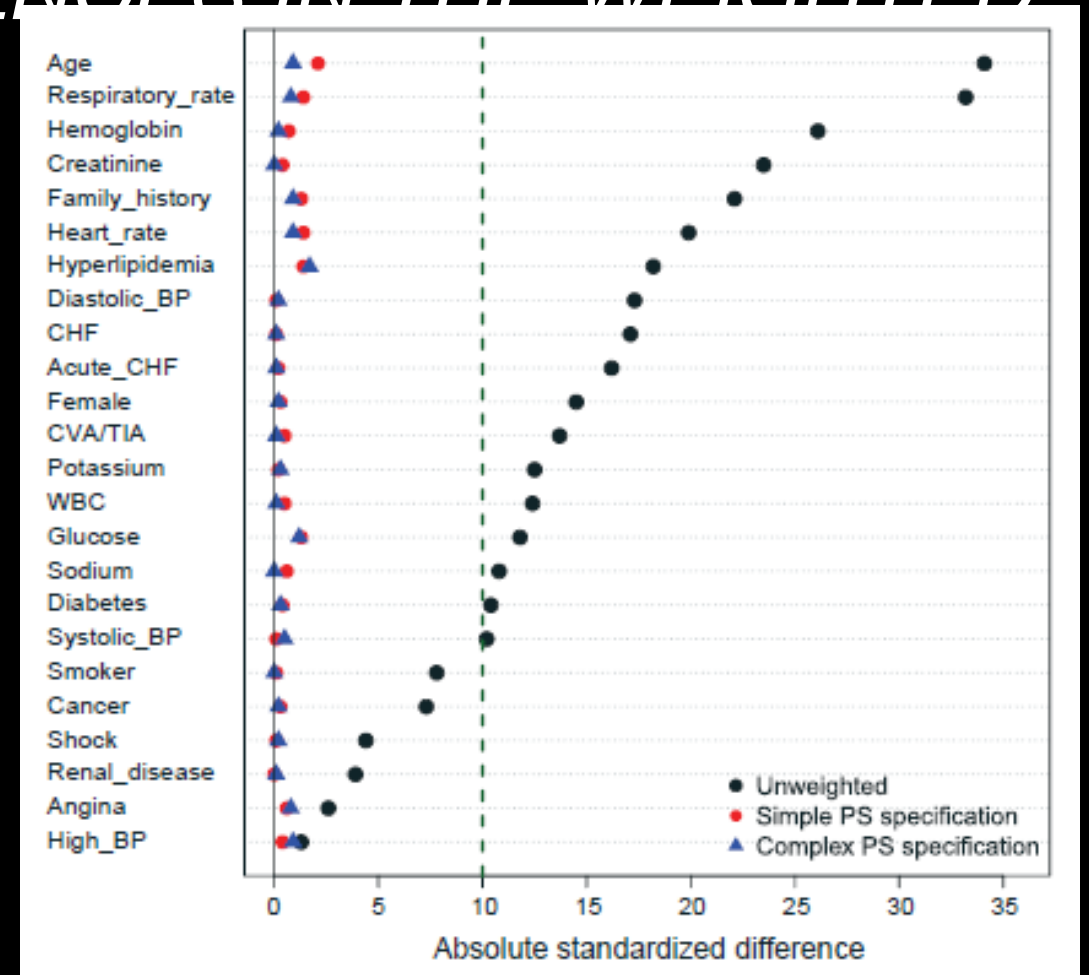
Table I. Baseline characteristics of treated and control subjects in original sample.

Variable	Beta-blocker: No (N=2929)	Beta-blocker: Yes (N=6178)	Standardized difference
<i>Demographic characteristics</i>			
Age	69.6 ± 13.5	65 ± 13.3	-34.1
Female	1144 (39.1%)	1984 (32.1%)	-14.5
<i>Presenting signs and symptoms</i>			
Cardiogenic shock	26 (0.9%)	32 (0.5%)	-4.4
Acute congestive heart failure (CHF)/pulmonary edema	214 (7.3%)	224 (3.6%)	-16.2
<i>Classic cardiac risk factors</i>			
Diabetes	842 (28.7%)	1494 (24.2%)	-10.4
Current smoker	916 (31.3%)	2158 (34.9%)	7.8
Hyperlipidemia	767 (26.2%)	2132 (34.5%)	18.2
Hypertension	1343 (45.9%)	2793 (45.2%)	-1.3
Family history of coronary artery disease	745 (25.4%)	2195 (35.5%)	22.1
<i>Comorbid conditions</i>			
Cerebrovascular disease/transient ischemic attack (CVA/TIA)	354 (12.1%)	493 (8%)	-13.7
Angina	975 (33.3%)	1982 (32.1%)	-2.6
Cancer	110 (3.8%)	154 (2.5%)	-7.3
Congestive heart failure (CHF)	189 (6.5%)	177 (2.9%)	-17.1
Renal disease	21 (0.7%)	26 (0.4%)	-3.9
<i>Vital signs on admission</i>			
Systolic blood pressure	146.8 ± 31.4	149.9 ± 30.9	10.2
Diastolic blood pressure	81.8 ± 18.6	84.9 ± 18.3	17.3
Heart rate	86.9 ± 25.9	82.1 ± 22.7	-19.9
Respiratory rate	22.2 ± 6.5	20.3 ± 4.8	-33.2
<i>Laboratory tests</i>			
Glucose	9.8 ± 5.2	9.2 ± 5.2	-11.8
White blood count	10.6 ± 5.5	10 ± 4.3	-12.4
Hemoglobin	135.2 ± 20	140.2 ± 17.7	26.1
Sodium	138.7 ± 4.2	139.2 ± 3.5	10.8
Potassium	4.1 ± 0.6	4.1 ± 0.5	-12.5
Creatinine	114.2 ± 77.4	98.8 ± 50.3	-23.5

Note: Continuous variables are represented as mean ± standard deviation, while dichotomous variables are represented as N (%).

IPTW BALANCE DIAGNOSTICS COMPARISON OF MEANS AND PREVALENCES IN THE WEIGHTED

- Simple specification model and Complex specification model
- The largest absolute standardized difference in the weighted sample was 2.1% (age) for simple model and 1.7% (hyperlipidemia) for complex model among the 24 baseline covariates.
- The standardized differences in the unweighted sample exceeded 10% for 18 (75%) of the 24 baseline covariates
- These diagnostic assessments suggest that weighting by the IPTW has created a sample in which the means of continuous baseline covariates and the prevalence of binary baseline variables are similar between treated and control subjects.



IPTW BALANCE DIAGNOSTICS

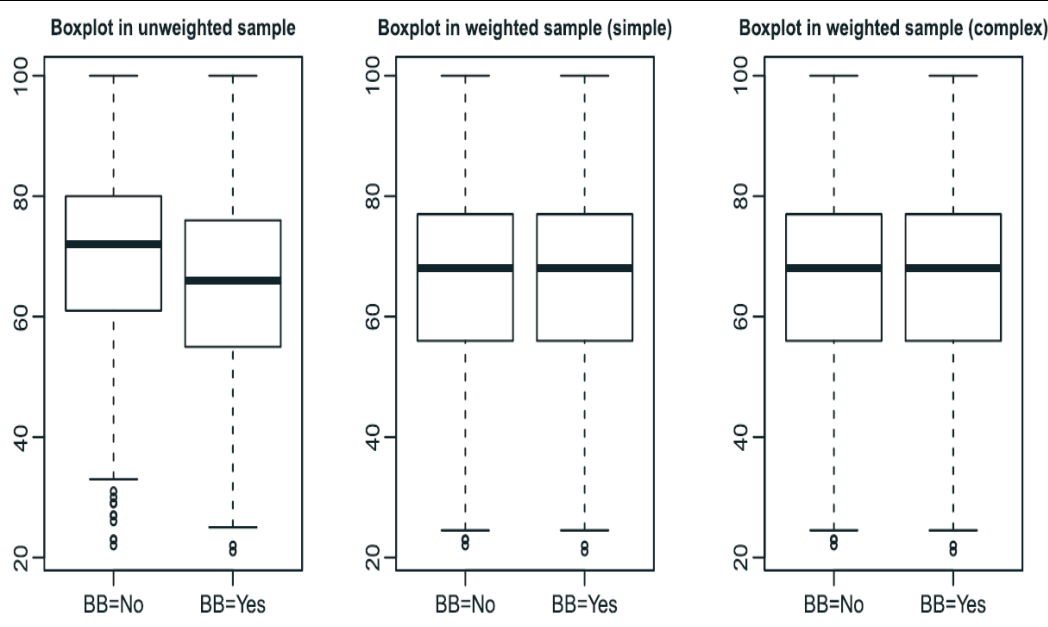
COMPARISON OF HIGHER-ORDER MOMENTS AND INTERACTIONS

Standardized differences	Unweight	Simple model	Complex model
Min	2.4%	0%	0%
P25	10.7%	0.5%	0.3%
P50	18%	1.0%	0.6%
P75	24.6%	1.5%	0.9%
Max	43.3%	2.3%	1.4%

- These analyses suggest that by weighting by the inverse probability of treatment, a sample has been created in which the means of higher-order terms and interactions between continuous variables are similar between treated and control subjects.
- Better balance was achieved using the weights derived from the complex specification of the propensity score model than using the simple specification; however, differences between the two approaches were at most modest.

IPTW balance diagnostics

Graphical comparisons of the distribution of continuous covariates

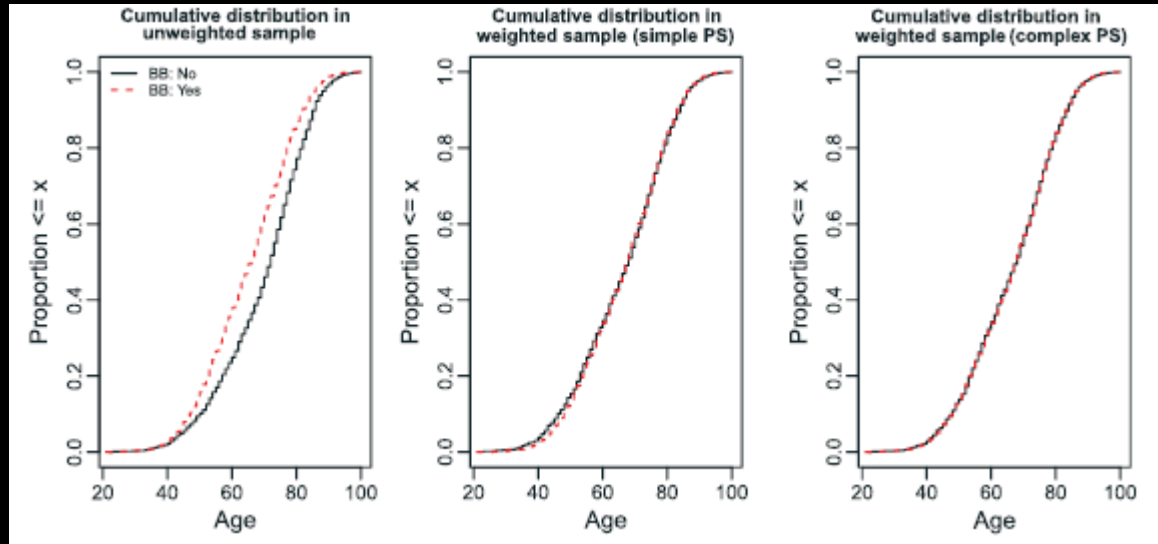


- The median age is greater in patients who did not receive a beta-blocker compared with patients who did receive a prescription for a beta-blocker at discharge.
- After weighting by the inverse probability of treatment, the two side-by-side boxplots appear nearly identical

Side-by-side box plots distribution of age between treated and control subjects.

IPTW balance diagnostics

Graphical comparisons of the distribution of continuous covariates



The empirical CDFs comparing the distribution of age between treated and control subjects.

- The distribution of age is shifted upwards in those who did not receive a prescription compared with those who did receive a prescription.
- Empirical cumulative distribution is nearly identical between treated and control subjects in both weighted samples.

IPTW balance diagnostics

Kolmogorov–Smirnov test statistic for comparing distribution of baseline covariates between treatment groups.

KS test	Unweight	Simple model	Complex model
Min	0.05	0.014	0.005
Max	0.164	0.027	0.02
Difference	0.114	0.013	0.015

The test statistic for each of the 11 variables was higher in the sample weighted by the simple specification of the propensity score than it was in the sample weighted by the complex specification of the propensity score.

INTERPRETATION OF BALANCE

- The interpretation of balance diagnostics is, to a certain extent, inherently subjective.
- The degree of imbalance that is acceptable likely depends on the magnitude of the effect of the covariate on the outcome.
- Thus, greater imbalance may be acceptable for covariates that are weakly prognostic than for covariates that are strongly prognostic.