Alternative Approaches For Confounding Adjustment In Observational Studies Using Weighting Based On The Propensity Score: A Primer For Practitioners

Pokket Sirisreetreerux

CEB Journal club

INTRODUCTION

- Propensity scores have become a cornerstone of confounding adjustment in observational studies evaluating outcomes of treatment use in routine care.
- 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.
- Can only achieve exchangeability with respect to the measured characteristics.

INTRODUCTION

- 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 including
  - matching
  - stratification
  - adjustment as a regressor
  - weighting

INTRODUCTION

- 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
- Method of choice for analysing observational data for many researchers

INTRODUCTION

- Matching
  - Discarding unmatched observations
  - Paradoxical phenomenon of increasing covariate imbalance

INTRODUCTION

- Weighting, Advantages including
  - Keeps most observations in the analysis so increase precision
  - Easily to transparent reporting of the balance achieved between treatment and reference populations
  - The most flexible approach of using propensity scores in the analysis with multiple available variations

INTRODUCTION

- Approaches for weighting
  - Traditional approaches
    - Inverse probability treatment weights (IPTW)
    - Standardised mortality ratio weights (SMRW)
  - Newer approaches
    - Propensity score fine stratification weights
    - Matching weights
    - Overlap weights

OBJECTIVE

- To demonstrate implementation and provide insights into the process of selecting a specific approach for a particular study
- Not to compare performance of different weighting methods

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.
- 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

- To account for the fact that the pseudo-population size is inflated or deflated relative to the original study population, a sandwich type estimator is recommended for variance estimation for the treatment effect estimates

TARGET OF INFERENCE (ESTIMAND)

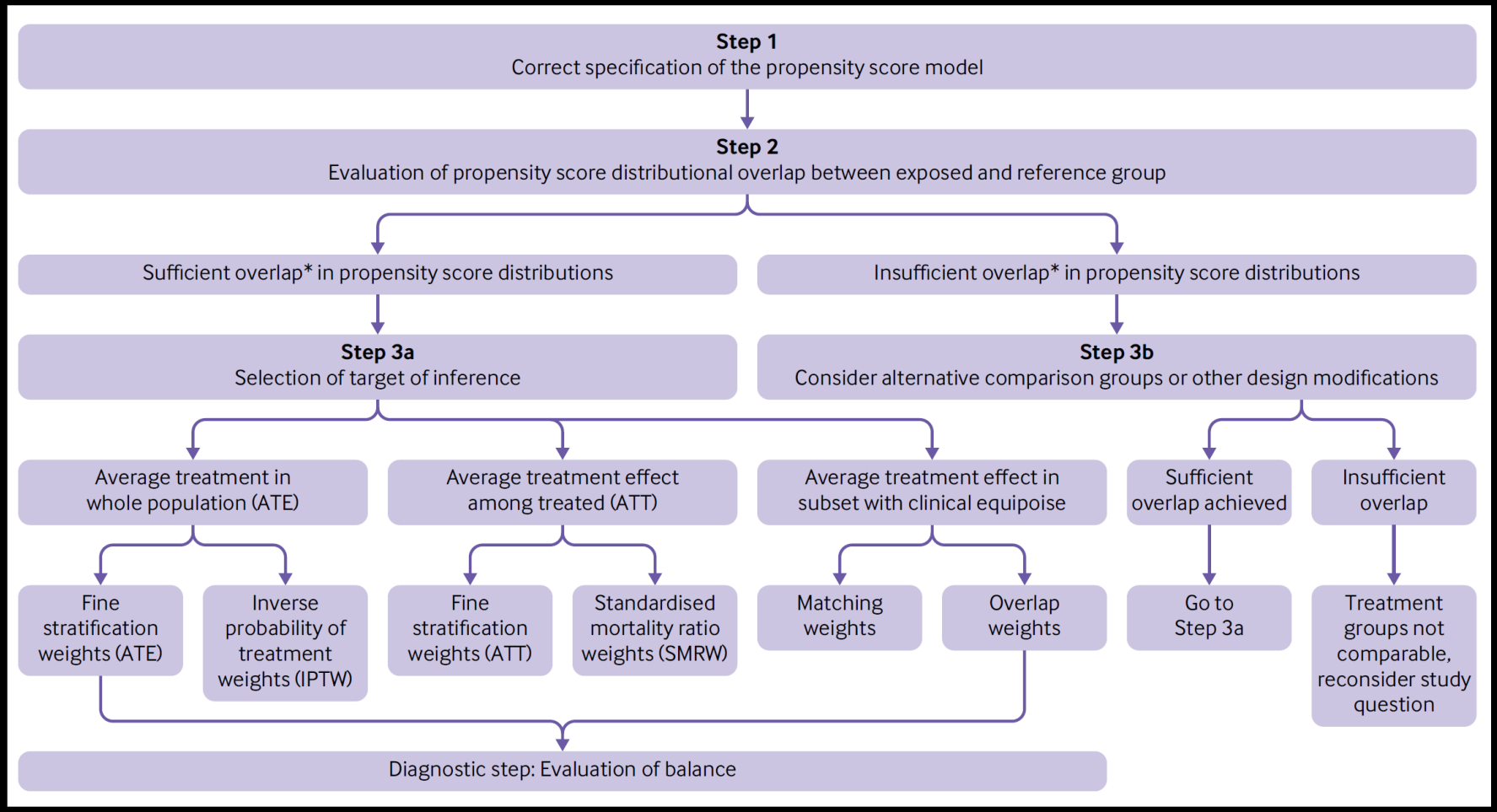
"Would it be feasible to treat all eligible patients included in the study with the treatment of interest?"

- If yes, the target of inference defined as the average treatment effect (ATE).
- If no, the target of inference might be defined as average treatment effect among the treated population (ATT).

TARGET OF INFERENCE

- In the absence of treatment effect heterogeneity by patient characteristics, ATE and ATT will coincide.

CONSIDERATIONS WHEN SELECTING A PROPENSITY SCORE WEIGHTING METHOD



STEP 1: CORRECT SPECIFICATION OF THE PROPENSITY SCORE MODEL

- Avoiding misspecification of the propensity score model

STEP 1: CORRECT SPECIFICATION OF THE PROPENSITY SCORE MODEL

- To construct the propensity score model,
  - Inclusion of outcome risk factors in the model
  - Exclusion of strong predictors of treatment that are not associated with outcomes to avoid increased variance and amplification of bias due to unmeasured confounding

STEP 1: CORRECT SPECIFICATION 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
- Learning approaches such as neural networks—could provide alternatives that are less prone to misspecification

STEP 1: CORRECT SPECIFICATION OF THE PROPENSITY SCORE MODEL

X Approaches that use the score directly to create weights such as IPTW more prone to increased bias and variance from misspecification of the propensity score model

✓ Weighting approach based on propensity score stratification might be more robust against misspecification …. use the score only to create stratums and then use the counts of observations within each stratum to derive weights.

STEP 1: CORRECT SPECIFICATION OF THE PROPENSITY SCORE MODEL

- Diagnostic step of checking covariate balance between the treatment and reference populations

1. Standardised difference in prevalence (recommended)

2. overall measure of balance, such as the post weighting C statistic, where values closer to 0.5 would indicate achievement of balance

STEP 1: CORRECT SPECIFICATION OF THE PROPENSITY SCORE MODEL

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.

STEP 1: CORRECT SPECIFICATION OF THE PROPENSITY SCORE MODEL

- Example

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)	20350 (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	19717 (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.

STEP 1: CORRECT SPECIFICATION OF THE PROPENSITY SCORE MODEL

- Example

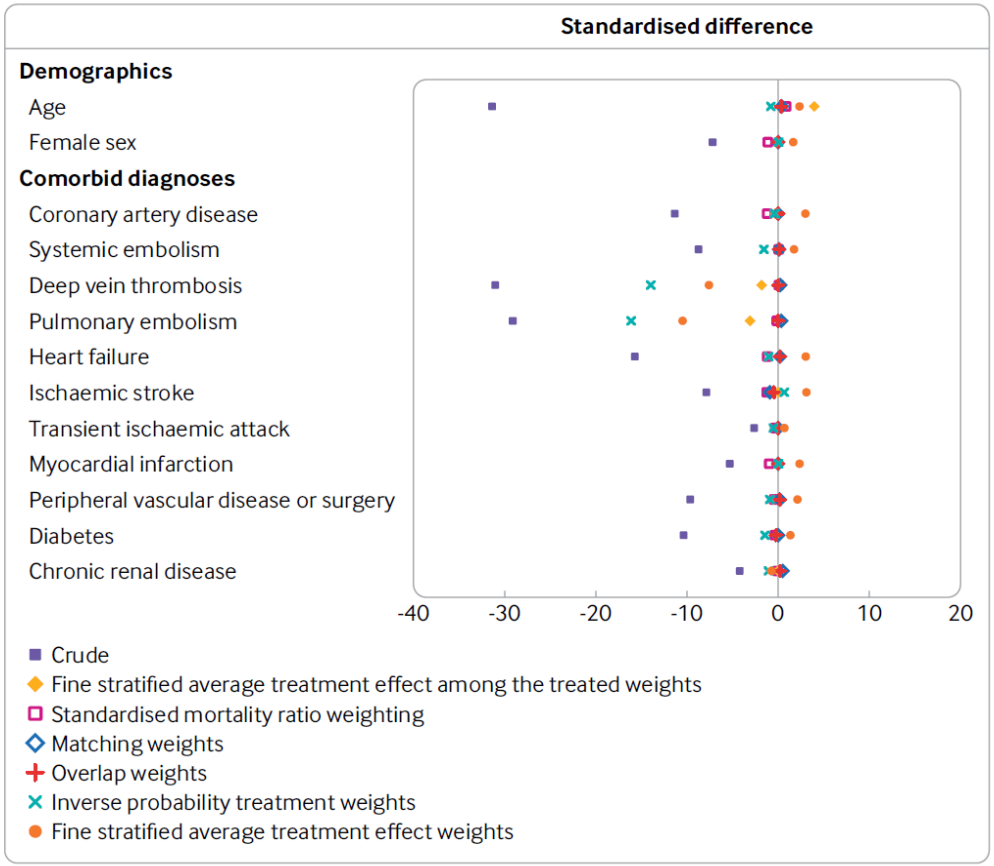
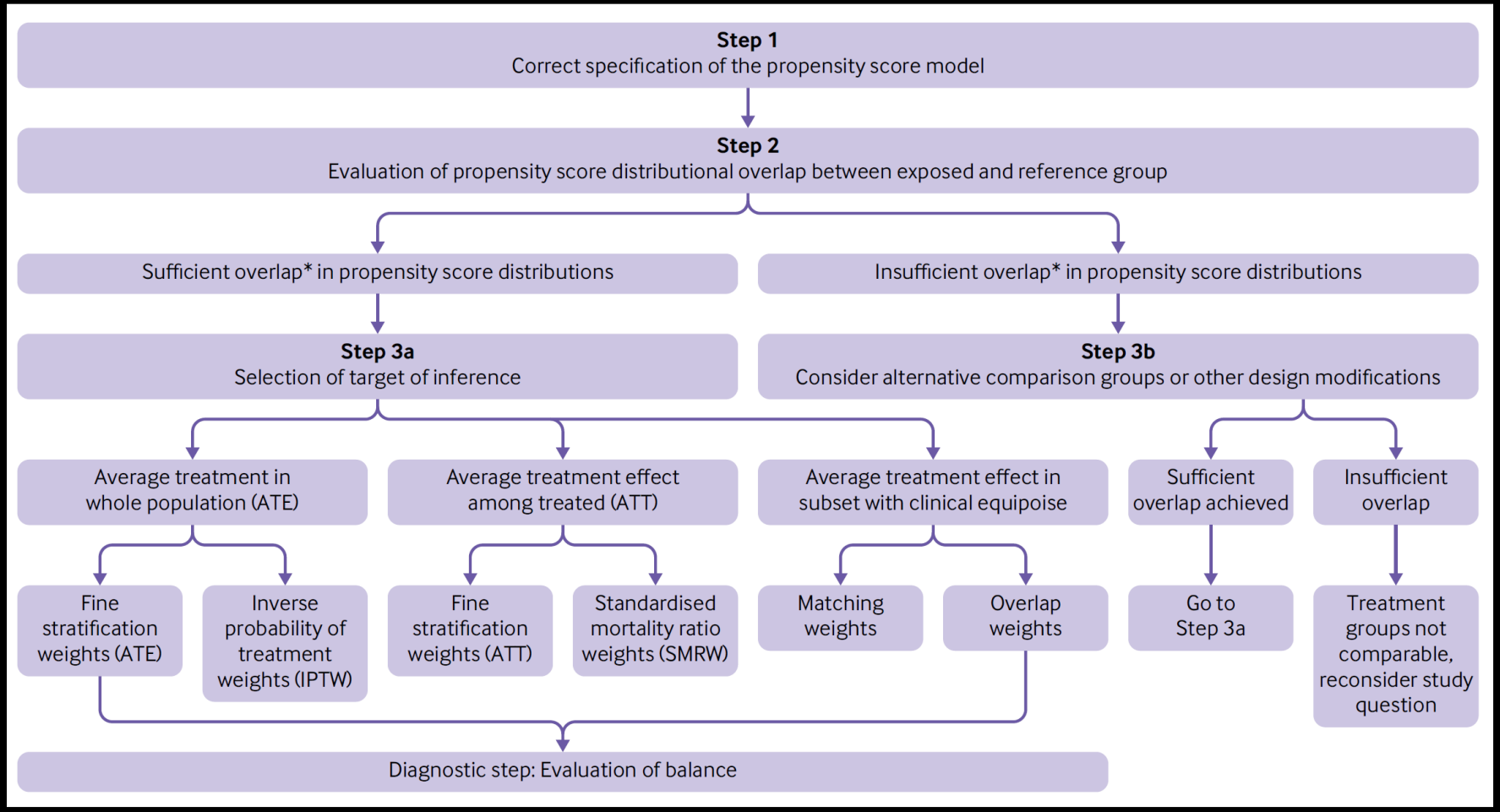


Fig 2 | Standardised differences before and after propensity score weighting, in case example of dabigatran versus warfarin initiation for atrial fibrillation, by selected patient characteristics

CONSIDERATIONS WHEN SELECTING A PROPENSITY SCORE WEIGHTING METHOD



STEP 2: EVALUATION OF PROPENSITY SCORE DISTRIBUTIONAL OVERLAP BETWEEN EXPOSED AND REFERENCE GROUPS

- High overlap generally indicates a reasonable degree of clinical equipoise in treatment selection.
- The general recommendation of trimming the regions of non-overlap to ensure restriction to regions where patients had a nonzero probability of receiving either treatment is especially important.
- 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.

STEP 2: EVALUATION OF PROPENSITY SCORE DISTRIBUTIONAL OVERLAP BETWEEN EXPOSED AND REFERENCE GROUPS

- 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.

- Example

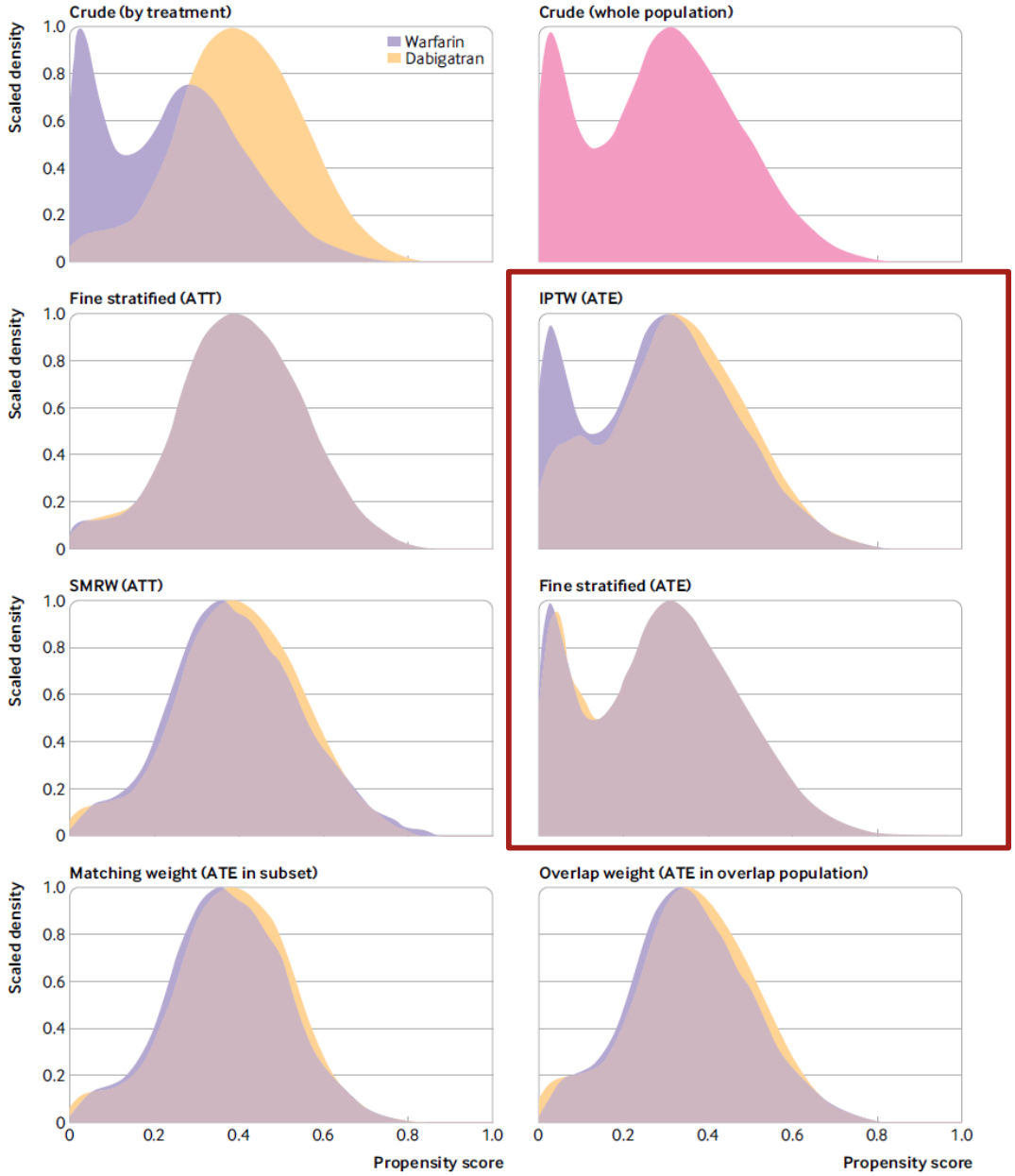
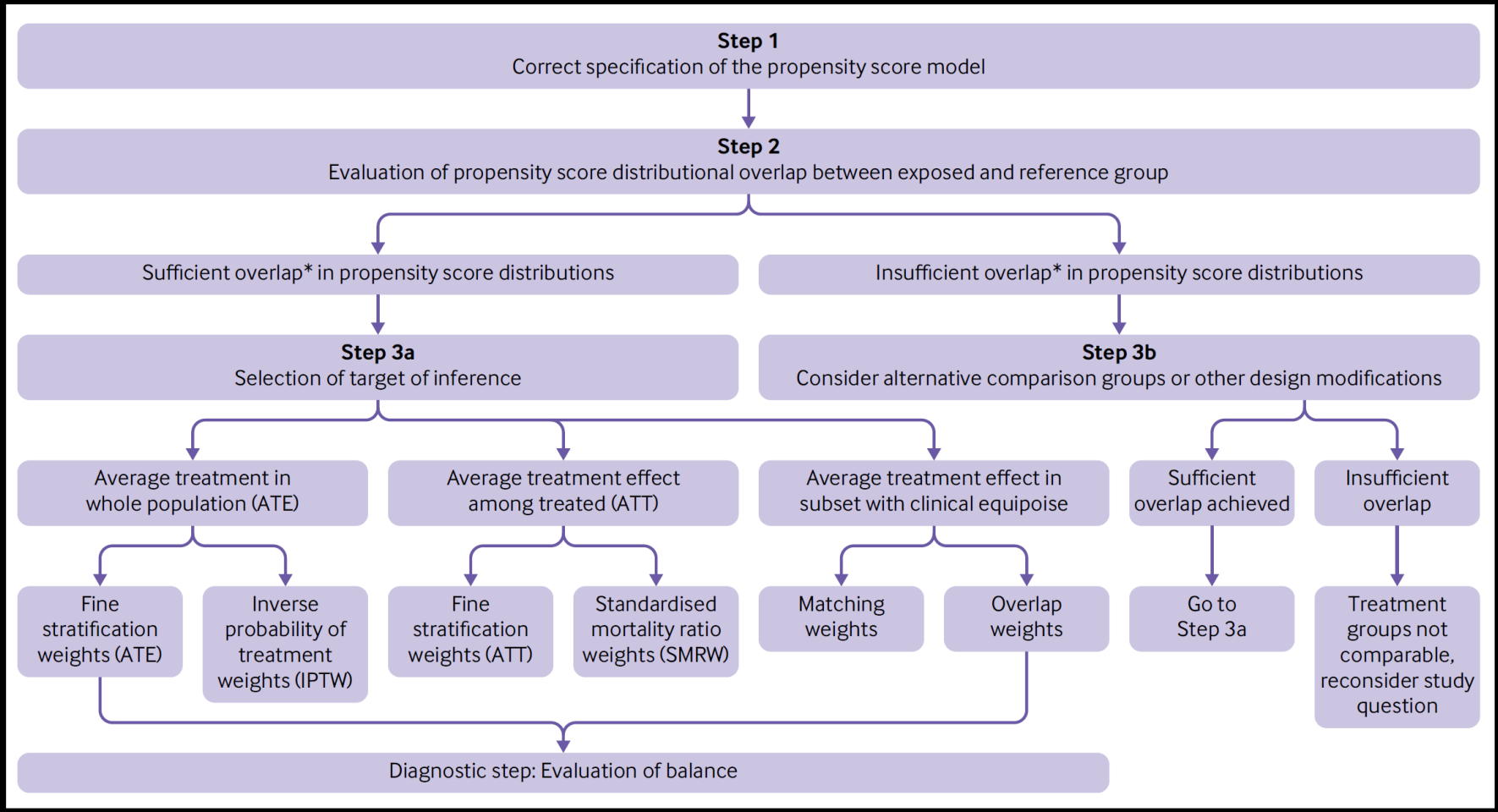


Fig 3 | Propensity score distributional overlap before and after propensity score weighting, in case example of dabigatran versus warfarin initiation for atrial fibrillation. ATE=average treatment effect; ATT=average treatment effect among the treated population; IPTW=inverse probability treatment weights;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

CONSIDERATIONS WHEN SELECTING A PROPENSITY SCORE WEIGHTING METHOD FOR CONFOUNDING ADJUSTMENT



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

1. ATE in the whole population
   - Two weighting approaches are available
   - 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.

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

Approach 1 Inverse probability treatment weighting (IPTW)

- weighting by the inverse probability of receiving the study treatment actually received

  treated group ; 1 / propensity score
  reference group ; 1 / (1−propensity score)

- Extreme weights are commonly observed whenever the propensity score is near 0 for a treated patient or near 1 for a reference patient.

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

- Approach 1 Inverse probability treatment weighting (IPTW)
  - 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.
  - Stabilizing weight, to prevent extreme weights by incorporating the marginal probability of receiving the treatment actually received in the numerator.

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

- Approach 1 Inverse probability treatment weighting (IPTW)
  - A special setting where IPTW is routinely used is in marginal structural modelling.
  - 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.

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

- Approach 2 Fine stratification weights targeting the average treatment effect (ATE)
  - not use the propensity score directly to calculate weights; instead, propensity scores are used to create fine stratums.
  - 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).

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

- Approach 2 Fine stratification weights targeting the average treatment effect (ATE)
  - Then, 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.

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

- Approach 2 Fine stratification weights targeting the average treatment effect (ATE)
  - 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.
  - mathematically equivalent to marginal mean weights

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

2. Average treatment effect among the treated population (ATT)

- Two weighting approaches are available
- aim to make the distribution of covariates in the reference group similar to the distribution observed in the treatment group.

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

Approach 1 Standardised mortality ratio weighting (SMRW)

- Setting weights to 1 for the treated patients and weighting reference patients by the odds of treatment probability: (propensity score/(1−propensity score)).
- Potentially vulnerable to extreme weights
- Weight truncation could be considered if large weights are observed.

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

Approach 2 Fine stratification weights targeting the average treatment effect among the treated population (ATT)

- Propensity scores are used to create fine stratums.
- 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
- Reference patients contribute proportionally to the relative number of total patients within a stratum
- Extreme weights are uncommon

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

3. ATE in a subset with clinical equipoise

- These approaches target the ATE in a subset of the overall population with some clinical equipoise.

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

3. ATE in a subset with clinical equipoise

- Approach 1 Matching weights
  - involves weighting patients based on a ratio of the lower of the two predicted probabilities to the predicted probability of the actually received treatment
  - extreme weights are impossible, no need for weight truncation.

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

3. ATE in a subset with clinical equipoise

Approach 1 Matching weights

- 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.

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

3. ATE in a subset with clinical equipoise

Approach 2 Overlap weights

- Weighting patients based on the predicted probability of receiving the opposite treatment
- Extreme weights are impossible, no need for weight truncation.
- This weighting method yields exact covariate balance between treated and reference groups by construction.

STEP 3B: CONSIDER ALTERNATIVE COMPARISON GROUPS OR OTHER DESIGN MODIFICATIONS (IF INSUFFICIENT OVERLAP)

- 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.

STEP 3B: CONSIDER ALTERNATIVE COMPARISON GROUPS OR OTHER DESIGN MODIFICATIONS (IF INSUFFICIENT OVERLAP)

- If sufficient overlap is achieved after modifications, then use of weighting based on the propensity score could be considered.
- If alternative comparison groups or design modifications fail to achieve sufficient overlap, investigators might need to reconsider the study question.

PROPENSITY SCORE BASED WEIGHTING APPROACHES OF OUTCOMES IN MORE THAN TWO TREATMENT 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.
- Generating propensity scores for three or more treatments in a multinomial logistic regression model.

PROPENSITY SCORE BASED WEIGHTING APPROACHES OF OUTCOMES IN MORE THAN TWO TREATMENT GROUPS

- For matching weights, the numerator includes the minimum of all available propensity scores for each patient and the denominator includes propensity of the treatment actually received.

PROPENSITY SCORE BASED WEIGHTING APPROACHES OF OUTCOMES IN MORE THAN TWO TREATMENT 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.

PROPENSITY SCORE BASED WEIGHTING APPROACHES OF OUTCOMES IN MORE THAN TWO TREATMENT GROUPS

- An extension of overlap weights, termed as generalised overlap weights, has been proposed for settings of three or more groups.
- Weights are constructed as the product of the inverse probability weights and the harmonic mean of the generalised propensity scores.
- These weights target the population with the most overlap in covariates across the multiple treatments.

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)	$(N_{\text{total in PS stratum } i}/N_{\text{total}}) / (N_{\text{exposed in PS stratum } i}/N_{\text{total exposed}})$	$(N_{\text{total in PS stratum } i}/N_{\text{total}}) / (N_{\text{reference in PS stratum } i}/N_{\text{total reference}})$	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.

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.

THANK YOU