Review

For reprint orders, please contact: reprints@futuremedicine.com

Journal of **Comparative Effectiveness Research**

Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants

Victoria Allan*[,1], Sreeram V Ramagopalan[1], Jack Mardekian[2], Aaron Jenkins[3], Xiaoyan Li[4], Xianying Pan[5] & Xuemei Luo[6]

[1]Centre for Observational Research & Data Sciences, Bristol-Myers Squibb, Uxbridge, UK
[2]Statistics, Global Biometrics and Data Management, Pfizer Inc., New York City, NY, USA
[3]Patient Health & Impact, Outcomes & Evidence, Pfizer Ltd, Tadworth, UK
[4]Worldwide Health Economics and Outcomes Research, Bristol-Myers Squibb, Lawrenceville, NJ, USA
[5]Pharmacoepidemiology, Bristol-Myers Squibb, Lawrenceville, NJ, USA
[6]Patient Health & Impact, Outcomes & Evidence, Pfizer Inc., Groton, CT, USA
*Author for correspondence: Tel.: +44 189 552 3382; victoria.allan@bms.com

After decades of warfarin being the only oral anticoagulant (OAC) widely available for stroke prevention in atrial fibrillation, four direct OACs (apixaban, dabigatran, edoxaban and rivaroxaban) were approved after demonstrating noninferior efficacy and safety versus warfarin in randomized controlled trials. Comparative effectiveness research of OACs based on real-world data provides complementary information to randomized controlled trials. Propensity score matching and inverse probability of treatment weighting are increasingly popular methods used to address confounding by indication potentially arising in comparative effectiveness research due to a lack of randomization in treatment assignment. This review describes the fundamentals of propensity score matching and inverse probability of treatment weighting, appraises differences between them and presents applied examples to elevate understanding of these methods within the atrial fibrillation field.

First draft submitted: 27 January 2020; Accepted for publication: 3 March 2020; Published online: 18 March 2020

Keywords: atrial fibrillation • comparative effectiveness research • confounding by indication • inverse probability of treatment weighting • oral anticoagulants • propensity score matching

Atrial fibrillation (AF) is a cardiac arrhythmia affecting millions worldwide. In 2010, the global prevalence of AF was estimated at 33.5 million [1] and projections indicate a doubling in the number of patients by 2050 [2], with the ageing population a key contributing factor. AF significantly increases the risk of thromboembolic stroke with patients five-times more likely to experience stroke compared with those without AF [3]. Oral anticoagulants (OACs) are effective in reducing the risk of stroke by 64%, according to a meta-analysis [4]. After decades of warfarin being the only OAC widely available, four direct OACs (DOACs; apixaban, dabigatran, edoxaban and rivaroxaban) were approved after demonstrating noninferior efficacy and safety versus warfarin in randomized controlled trials (RCTs) [5–8].

RCTs evaluate the efficacy and safety of OACs in well-controlled environments with precisely defined study population inclusion and exclusion criteria. RCTs are purposefully designed to minimize sources of bias and provide the ideal conditions for testing whether a cause–effect relationship exists between a treatment and an outcome [9]. Real-world evidence (RWE), derived from data collected in the normal delivery of clinical care, provides complementary information to that obtained from RCTs and is of increasing interest to healthcare decision makers [10]. RWE can be used to establish whether trial efficacy translates into real-world effectiveness, or to explore how a treatment performs in more diverse patient populations with differing levels of adherence

Future Medicine

and compliance, over a longer period of follow-up and in comparative effectiveness research (CER) versus other therapies that have not been examined under trial conditions [11].

The introduction of DOACs has significantly expanded the number of treatment options available to AF patients, yet to date there have been no RCTs comparing the efficacy of one DOAC against another. In the absence of head-to-head RCTs, the comparative effectiveness of DOACs continues to be extensively researched using RWE studies. In a recent systematic review and meta-analysis, Li and colleagues identified 15 RWE studies comparing rivaroxaban versus dabigatran, nine studies comparing rivaroxaban versus apixaban and nine studies comparing apixaban versus dabigatran; there were no comparisons with edoxaban, reflecting its later approval date. The meta-analyzed results showed no significant differences between the three DOACs compared in terms of stroke or systemic embolism risk, whereas apixaban was associated with a significantly lower risk of major bleeding compared with rivaroxaban and dabigatran, and rivaroxaban conferred a significantly higher risk of major bleeding compared with dabigatran. Li and colleagues concluded that the study results may help to guide treatment decisions on the choice of DOACs for patients with AF. However, highlighting a key limitation, the authors reflected that DOAC treatments were not prescribed at random, therefore there could be confounding by indication bias influencing the findings. The potential for bias among the included studies was assessed using the Cochrane Collaboration ROBINS-I (Risk of Bias In Non-randomized Studies of Interventions) evaluation tool, among which 'confounding' and 'selection of participants' are two key assessment domains [12,13].

Confounding by indication may arise in real-world comparative effectiveness studies when there are inherent differences in the patients prescribed the two or more treatments being compared. As defined by Joseph and colleagues, confounding by indication refers to bias in relationship between a treatment and the intended outcome of the treatment due to the clinical reasons for the treatment. The indication for treatment is based upon both the physician's and the patient's judgement of the disease severity, prognosis and expected therapeutic effect of the treatment [14]. RCTs remain the gold standard study design for evaluating the relative benefits and/or risks of treatments, because the process of randomization minimizes imbalances in both observed and unobserved factors that could introduce bias into the assignment of patients to the treatment groups being compared [15]. In an RCT comparing two treatments, patients are equally likely (i.e., have a 50% chance) to receive either treatment, therefore this maximizes the probability that any resulting differences in outcomes between groups are truly due to differences in the treatments and are not due to any observed or unobserved differences in the patients receiving each treatment. CER using real-world data therefore requires rigorous statistical methods, often extending beyond multivariable regression models, to account for the lack of randomization in treatment assignment and control for confounding by indication [16]. Unlike an RCT, these methods are however restricted to addressing only observed confounders (i.e., factors that are measured and collected in the study).

Propensity score matching (PSM) and inverse probability of treatment weighting (IPTW) are increasingly popular methods used to address confounding by indication in RWE studies. Within the AF field, the number of research publications referencing these methods has been increasing year-on-year, as indexed in the PubMed database (Figure 1). Researchers have traditionally relied upon multivariable regression models to adjust for differences in patient characteristics, however, they are now turning to PSM and IPTW because of the stronger theoretical and statistical basis for these methods that has been argued in the literature [17]. While, for the most part, multivariable regression and propensity score-based methods (such as PSM and IPTW) have been found to lead to similar study conclusions, [18,19] from a theoretical standpoint, PSM and IPTW aim to achieve a balanced distribution of confounders across treatment groups and thereby more closely emulate the properties of an RCT [20]. From a statistical standpoint, propensity score-based methods have been shown to lead to more robust and less biased estimations of the treatment effect when there are few outcome events relative to the number of potential confounders (i.e., fewer than eight events per confounder) [21]. In Li and colleagues' systematic review of 15 real-world studies reporting comparisons between DOACs, two thirds of studies continued to use multivariable logistic or survival regression models alone to address confounding, with one third of studies opting for more robust PSM and IPTW methods. The studies using PSM and IPTW methods were more often assessed as being at low risk of bias [12].

While PSM and IPTW endeavor to achieve the same objective in balancing out differences between treatment groups, the two methods provide a different measurement of the treatment effect and this should be interpreted accordingly. When applied to the same data, PSM and IPTW may not always point to the same findings suggesting that these methods are not strictly interchangeable. For example, a study comparing dabigatran versus warfarin among real-world AF patients reported a hazard ratio of 0.77 (95% confidence interval: 0.54–1.09) for the risk of stroke and 0.75 (0.65–0.87) for the risk of major bleeding when estimated using PSM, contrasting with 0.00 (0.00–

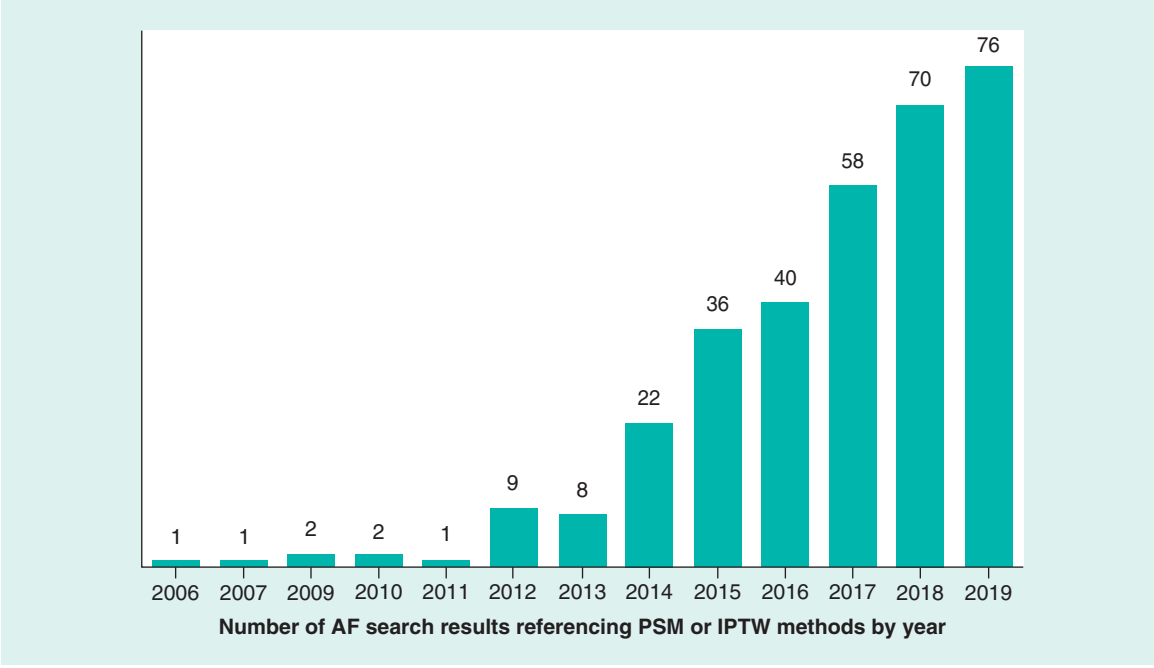

Figure 1. Increase in propensity score matching and inverse probability of treatment weighting methods within comparative effectiveness research of oral anticoagulants (2006–2019). Search results per year were downloaded from the PubMed database (https://www.ncbi.nlm.nih.gov/pubmed/) on 29-Nov-2019 based on the following search terms: (anticoagulant* OR warfarin OR apixaban OR dabigatran OR rivaroxaban OR edoxaban) AND (atrial fibrillation) AND (propensity score OR inverse probability of treatment weighting).
AF: Atrial fibrillation; IPTW: Inverse probability of treatment weighting; PSM: Propensity score matching.

0.56) and 0.08 (0.08–0.10), respectively, when estimated using IPTW [22]. As shown, PSM and IPTW procedures yielded substantially different estimates of the relative risk reduction in bleeding associated with dabigatran and conflicting findings with regards to stroke. In subsequent analyses re-calculating the IPTW results, a hazard ratio of 0.57 (0.46–0.71) for the risk of stroke and 0.75 (0.69–0.82) for the risk of major bleeding was obtained, which were closer (but not identical) estimates to the PSM results. Therefore, as these methods become more widely used in real-world studies comparing the effectiveness and safety of OACs, there is an important need to improve understanding of the fundamentals of PSM and IPTW in order to cast a critical eye over study findings within the field of AF research.

To this end, this review describes PSM and IPTW methods to address confounding by indication in real-world studies, appraises similarities and differences between these techniques and alongside gives illustrative examples from some case studies comparing the effectiveness and safety of OACs. A clearer conceptualization of PSM and IPTW methods will help to ensure the findings of comparative studies are correctly interpreted, communicated and implemented into AF-related treatment decisions in clinical practice.

Case studies

Four case studies involving the use of PSM, IPTW or a combination of both methods in real-world CER of OACs will be presented throughout this article to demonstrate how these methods have been applied in practical terms. These case studies were selected from a range of researchers, data sources and geographies (Table 1). The ARISTOPHANES study from USA [23] and the SAKURA AF Registry study from Japan [24] used PSM to compare the effectiveness and safety of apixaban, dabigatran, rivaroxaban and warfarin among nonvalvular AF patients [23]. Using the Danish nationwide databases, Larsen and colleagues, [25] instead used the IPTW method to investigate the comparative effectiveness and safety of the same four OACs. Seeger and colleagues, [22] used both PSM and IPTW techniques to compare warfarin with dabigatran in two commercial health insurance databases in the USA.

10.2217/cer-2020-0013

Table 1. Selected case studies of real-world comparative effectiveness research of oral anticoagulants using propensity score matching and/or inverse probability of treatment weighting.

Study (year)	Data source	OACs	Main study end points	Variables included in propensity score	Statistical analysis	Ref.
Lip (2018)	ARISTOPHANES, USA	Warfarin, dabigatran, rivaroxaban, apixaban	– Stroke/SE, including ischemic stroke, hemorrhagic stroke and SE – Major bleeding, including gastrointestinal bleeding and intracranial hemorrhage	Age, gender, US region, Charlson Comorbidity Index, bleeding history, congestive heart failure, diabetes mellitus, hypertension, renal disease, liver disease, myocardial infarction, dyspepsia/stomach discomfort, peripheral vascular disease, stroke/SE, TIA, anemia/coagulation defects, alcoholism, concomitant use of: ACE/ARBs, amiodarone, beta-blockers, H2-receptor antagonists, proton pump inhibitors, statins, antiplatelet, NSAIDS	PSM: 1:1 propensity score matching. Nearest neighbor matching method without replacement with a caliper of 0.01, was used to match the patients	[23]
Okumura (2018)	SAKURA AF Registry, Japan	Warfarin, dabigatran, rivaroxaban, apixaban	– Stroke (ischemic stroke, hemorrhagic stroke or TIA) or SE – Major bleeding	Age, sex, height, weight, paroxysmal AF, hypertension, diabetes mellitus, heart failure, vascular disease, stroke or TIA, major bleeding, history of ablation, antiplatelet agent use, nonsteroidal anti-inflammatory drug use, OAC therapy duration and CrCl	PSM: matching was achieved by a 1:1 nearest neighbor approach (without replacement) within a caliper of 0.05	[24]
Larsen (2016)	Danish nationwide databases, Denmark	Warfarin, dabigatran, rivaroxaban, apixaban	– Ischemic stroke; a composite of ischemic stroke or systemic embolism; death; and a composite of ischemic stroke, systemic embolism or death – Any bleeding, intracranial bleeding, major bleeding	Age, sex; ischemic stroke or systemic embolism or transient ischemic attack; vascular disease; hypertension; diabetes; cancer; recent prescription of aspirin, beta-blockers, nonsteroidal anti-inflammatory drugs, or statins; and CHA2DS2-VASc and HAS-BLED scores	IPTW: weights calculated using generalized boosted models, based on 10,000 regression trees	[25,45]
Seeger (2016)	Two commercial health insurance databases (MarketScan® / Clinformatics®), USA	Warfarin, dabigatran	– Hospitalization for stroke (hemorrhagic or ischemic) – Hospitalization for major bleeding, including intracranial and extracranial bleeding	Age, sex, census region, calendar time, coronary artery disease, systemic embolism, DVT, PE, hypertension, diabetes, hyperlipidemia, atherosclerosis, heart failure, stroke, intracranial bleeding, hemorrhagic stroke, ischemic stroke, previous TIA, CHADS2 score, CHA2DS2-VASc score, previous MI, HAS-BLED, peptic ulcer disease, upper GI bleed, lower/unspecified GI bleed, urogenital bleed, other bleeds, peripheral vascular disease or PVD surgery, prior liver disease, prior cancer, renal dysfunction, acute renal disease, chronic renal insufficiency, diabetic nephropathy, hypertensive nephropathy, miscellaneous renal insufficiency, aspirin, aspirin/dipyridamole, clopidogrel, prasugrel, ticagrelor, other antiplatelet agents, NSAIDs, heparin, low-molecular weight heparins, PGP inhibitors, ARB, ACE inhibitor, Beta blocker, calcium channel blocker, other hypertension drugs, antiarrhythmic drugs (other than amiodarone and dronedarone), statin, other lipid-lowering drugs, diabetes medications, antidepressants, antipsychotics, anticonvulsants, proton pump inhibitor, H2 receptor antagonist, other gastroprotective agents, vitamin K therapy, number of medications, number of hospitalizations, number of hospital days, number of office visits, number of cardiologist visits, number of neurologist visits, hospitalization in 30 days prior to treatment initiation, number of laboratory tests ordered, number of INR (prothrombin) tests ordered, number of lipid tests ordered, number of creatinine tests ordered, treating prescriber	PSM: A 1:1 PS match was performed using the nearest neighbor methodology with a maximum caliper of 0.05 and additional matching within the calendar quarter to improve balance with respect to prescribing behaviors that might change over time IPTW: weighting was performed using IPTW (1/PS for dabigatran initiators and 1/(1−PS) for warfarin initiators). The IPTW analysis was repeated after capping (truncating) the weights at a value of 4.0.	[22]

ACE: Angiotensin-converting enzyme; AF: Atrial fibrillation; ARB: Angiotensin II Receptor Blockers; DVT: Deep vein thrombosis; INR: International normalised ratio; IPTW: Inverse probability of treatment weighting; NSAID: Nonsteroidal anti-inflammatory drugs; OAC: Oral anticoagulant; PE: Pulmonary embolism; PSM: Propensity score matching; SE: Systemic embolism; TIA: Transient ischemic attack.

Propensity scores: the basis for PSM & IPTW

Propensity scores form the basis for both PSM and IPTW methods, however as later described, the differentiating step is in how propensity scores are then used to control for differences in the characteristics of patients receiving the treatments being compared.

As defined by Rosenbaum and Rubin, an estimated propensity score reflects the probability of treatment assignment conditional on a patient's measured baseline characteristics, [26] such as their age, gender, comorbidities and concurrent medications. In the real-world setting, and in direct contrast to an RCT, treatments are not prescribed at random and are instead likely influenced by the characteristics of the patients at the time a treatment decision was made. For example, the decision to prescribe an OAC to an AF patient is likely informed by their stroke and bleeding risks. A patient's pretreatment characteristics can therefore be used to predict, and provide a measure (from 0 to 1) of, how likely the patient is to receive a treatment. Through leveling out the distribution of scores across treatment groups, propensity scores can minimize confounding by indication bias and provide a fairer comparison of different treatments [27].

In practice, propensity scores are estimated using regression-based methods and are easily implemented with statistical software packages such as SAS, Stata and R [28–30]. Most often, when comparing two treatments, a logistic regression model is used in which treatment assignment (a binary dependent variable with value 1 if the patient receives the treatment of interest and value 0 if the patient receive the comparator treatment) is regressed upon patient baseline characteristics (independent variables) [27]. When selecting variables to be included in the propensity score estimation, variables that are both related to the exposure (i.e., treatment assignment) and to the outcome (i.e., the main study end points) are recommended to provide the most precise estimates [31]. Misspecification of the propensity score can introduce bias and lead to invalid inferences therefore careful variable selection is essential. Extensions of the propensity score, such as the doubly robust method, facilitate more systematic variable selection and provide greater protection against model misspecification, however, are not yet widely known or adopted among researchers [32,33]. Three of the four case studies of this review used logistic regression models to generate the propensity scores, [22–24] whereas the fourth study used generalized boosted models, an advanced technique based on machine learning methods, applicable when comparing more than two treatments simultaneously [23]. Common variables included in the propensity score estimations were age, gender, bleeding history, stroke history, comorbidities and concomitant meditations (Table 1).

Once the propensity scores are estimated, a range of different ways can be used to balance the distribution of the scores, and in turn confounding factors, across treatment groups being compared. This includes propensity score adjustment, matching, stratification and weighting [27]. Among them, PSM and IPTW are emerging as the most commonly used propensity score methods within the field of AF research. Hence, the focus of this review is to provide foundational understanding of PSM and IPTW and highlight fundamental differences in methodology and interpretation of these two increasingly implemented techniques. Further resources on adjustment and stratification, and alternative weighting procedures to IPTW, are available elsewhere [27,34,35].

The main conceptual difference between PSM and IPTW relates to notion of the treatment effect being estimated. Whereas IPTW estimates the average treatment effect (ATE), PSM estimates the average treatment effect for the treated (ATT). The ATE reflects the effect of the treatment in the scenario that every patient within the population was offered the treatment. In essence, this shifts the entire population from untreated to treated. This contrasts with the ATT, which reflects the effect of the treatment only among those who were ultimately treated [27]. The research question being addressed will guide the decision as to whether an estimation of the ATE or ATT is of greater relevance. Estimation of the ATE may be less appropriate when a large proportion of patients are not good candidates for one of the treatment options. An example is the comparison of low dose apixaban patients with all warfarin patients, as some warfarin patients do not meet apixaban low dose criteria, it would be unrealistic to estimate the treatment effect among all patients [36]. Grasping the conceptual differences between the ATE and ATT is crucial to the correct interpretation of results from the application of PSM and IPTW.

PSM: estimating the ATT

PSM works by matching patients receiving the treatment of interest with patients receiving the comparator treatment based on the estimated value of their propensity score. PSM provides an estimation of the ATT, because only patients who received the treatment of interest are matched with comparators for comparative analyses [27].

In practice, patients are matched with others who have a similar propensity score value, as an exact score is not always possible. A variety of different matching methods are available [37] with four main analytics decisions:

10.2217/cer-2020-0013

Table 2. An overview of available propensity score matching techniques to pair up patients receiving the treatment of interest with suitable comparator patients.

Matching techniques	Definition	Considerations
Matching patients one-to-one (1:1) or many-to-one (m:1):		
1:1 matching	Matching patients 1:1 means that only one patient in the comparator group is selected for each treated patient	– 1:1 matching is most commonly used – m:1 has the advantage of retaining more patients in the analysis, however, the quality of second or third choice matches may be lower which increases bias
m:1 matching	Matching patients m:1 means that multiple comparator patients are selected for each treated patient	
Matching patients with or without replacement:		
Matching without replacement	Matching without replacement means that once a patient from the comparator group has been matched, they cannot be selected as a comparator for another treated patient	– Matching without replacement is more commonly used – Matching with replacement could be beneficial if the number of comparator patients is small, but may mean some comparators are matched multiple times and others not at all
Matching with replacement	Matching with replacement allows for a patient in the comparator group to be matched and act as a comparator for more than one of the treated patients	
Matching patients using a greedy or optimal technique:		
Greedy nearest neighbor matching	Using a greedy matching technique means that treated patients are randomly selected one at a time to be matched to their nearest comparator from the pool of remaining comparators still available to be matched. Here the closest match is made irrespective of whether the selected comparator would have been a better match for another treated patient	– Greedy nearest neighbor matching is more commonly used – Greedy and optimal matching in general perform similarly well in creating balanced treatment and comparator groups – Optimal matching may be preferred when the sample of comparator patients is small
Optimal matching	With optimal matching, the goal is to pair up treated and comparator patients such that the best possible combination of nearest matches are made	
Matching patients with a caliper width:		
Without a caliper width	When propensity score matching is performed without a caliper width, this means that matches are made without any limit on the distance between the propensity score values in each pair of matches	A caliper width is recommended to match patients within a given threshold. Without a caliper threshold, treated patients could be matched to their nearest comparator even if the differences in propensity scores were very large
Applying a caliper width	The role of a caliper width is to set a maximum distance between the propensity score values in each pair of matches	

matching patients one-to-one (1:1) or many-to-one (m:1) [38]; matching patients with or without replacement [39]; matching patients using a greedy or optimal technique [39]; and whether to apply a caliper width (Table 2) [40].

First, matching patients 1:1 means that only one patient in the comparator group is selected for each treated patient, as opposed to selecting multiple comparator patients in m:1 matching. The rationale for selecting more than one comparator is to make more use of the available data by retaining more patients in the analysis, which in turn may also lead to increased precision. However, matching patients m:1 has also been shown to increase bias because second or third choice matches may be of lower quality [39]. Matching patients 1:1 is most commonly used in practice [38,41].

Second, matching without replacement means that once a patient from the comparator group has been matched, they cannot be selected as a comparator for another treated patient. Conversely, matching with replacement allows for a patient in the comparator group to be matched and act as a comparator for more than one of the treated patients. Matching patients with replacement serves to increase the quality of matching and could be useful if there are few comparator patients relative to number of patients receiving the treatment of interest. However, with this approach it is possible that a single comparator patient could be matched multiple times, whereas another potential comparator may not be matched at all. This could result in the treatment effect being estimated based on a very restricted subset of comparators. Furthermore, when matching with replacement, the matched pairs are no longer independent and this must be accounted for in the subsequent analysis and interpretation of study findings [39]. Matching patients with replacements has seldom been implemented in practice [37].

Third, using a greedy matching technique means that treated patients are randomly selected one at a time to be matched to their nearest comparator from the pool of remaining comparators still available to be matched. Here the closest match is made irrespective of whether the selected comparator would have been a better match for another treated patient. In contrast, optimal matching intends to pair up treated and comparator patients such that the best possible combination of nearest matches are made. Optimal matching may be preferred over greedy matching when the sample of comparator patients is small, however when there is less competition for matches, optimal matching does not usually perform any better than greedy matching in creating balanced treatment and comparator groups. Greedy matching is more commonly used in practice [37,39].

Fourth, the role of a caliper width is to set a maximum distance between the propensity score values in each pair of matches. Without a caliper threshold, treated patients could be matched to their nearest comparator even if the differences in propensity scores were very large. In studies to date, there has been a lack of consistency in the maximum distance selected for the caliper width, however recent simulations support using a caliper width equal to 0.2 of the standard deviation of the logit of the propensity score [40].

Three out of the four case studies included this review implemented PSM techniques in their CER of OACs (Table 1) [22–24]. In the ARISTOPHANES study, Lip and colleagues used 1:1 nearest neighbor matching without replacement with a caliper width of 0.01 [23]. In the SAKURA AF Registry study, Okumura and colleagues also used 1:1 nearest neighbor matching without replacement, however, with a caliper width set at 0.05 of the standard deviation of the logit of the propensity score [24]. Similarly, in their analysis of two commercial health insurance databases in USA, Seeger and colleagues used 1:1 nearest neighbor matching with a caliper width of 0.05 and additional matching on calendar quarter to account for any changes in prescribing behaviors over time. It was not reported whether patients were matched with or without replacement [22]. The main difference between the three studies in terms of their chosen PSM approach is in the selection of the caliper width. The selection of a narrow caliper distance ensures patients are more closely matched on the value of the propensity score, however selecting a distance that is too narrow could result in too few matches being made [40].

Once the matched cohort of treated and comparator patients has been formed, the fundamental next step is to verify that a balanced distribution of patient characteristics across treated and comparator patients has been achieved by way of using the propensity score. This check could be made using statistical significance tests (e.g. X-square/*t*-test), however, computing the standardized differences of each baseline variable is now a more common practice. Lip and colleagues, and Seeger and colleagues both reported computing standardized differences with a threshold of 10% to check whether adequate balance of covariates across treatment and control groups had been achieved through the PSM process [22,23].

Were any systematic differences to be found between the treatment and comparator cohorts, the propensity score model may require some modifications on the included variables and/or matching procedure. Alternatively, subsequent adjustments could be made in the main analysis to account for any remaining imbalances in patient characteristics. Variables related to both the outcome and treatment assignment have been shown to provide the best balance of patient characteristics across treatment groups [31]. However, a common misperception for the development of the propensity score is to aim for perfect prediction of treatment assignment and the inclusion of variables that are related only to treatment assignment should be avoided [42]. Researchers should set out *a priori* which approach will be used in the event that balance in patient characteristics is not achieved through PSM (i.e., whether the propensity score will be re-estimated or if post-PSM adjustments will be made). Once satisfied with the PSM step, the comparative assessment can be performed, for example, comparing the risk of the study main outcome with hazard ratios from Cox proportional hazards regression models or some other analytic method of choice. What is advantageous about PSM is that the method matches patients who truly received the treatment with those who truly received the comparator to allow a direct comparison of the outcomes of these two groups. The method is transparent, readily understood and easy to communicate. One frequently discussed drawback of the PSM method is that patients who could not be matched are as a result excluded from the analysis. Unmatched patients may be systematically different from matched patients, limiting the representativeness of the study population and generalizability of the overall study findings. Yet, an alternative viewpoint is that this form of study population restriction ensures overlap across treatment and comparator groups on the most important patient characteristics thereby removing any individuals who are extreme outliers and arguably should not be compared [43]. Another consideration is that PSM is limited to pair-wise comparisons of one treatment against another. If more than one treatment comparison is being made, for example comparing apixaban versus warfarin and apixaban

10.2217/cer-2020-0013

versus dabigatran, the matched pairs for each of the comparisons are likely to be different, so drawing inferences across these two comparisons may not be appropriate.

IPTW: estimating the ATE

In the IPTW method, weights are assigned to patients based on the inverse of their probability of receiving treatment, as estimated by the propensity score. IPTW results in a pseudo-population in which patients with a high probability of receiving treatment have a smaller weight and patients with a low probability of receiving treatment have a larger weight and thus the distribution of measured patient characteristics used to calculate the propensity score becomes independent of treatment assignment. IPTW provides an estimation of the ATE, because the study population is re-weighted to assess the effects of the treatment in the scenario that it was offered to all patients within the population [27,44].

When comparing two treatments, the weight for each patient is calculated by inverting the probability of receiving the treatment the patient did in fact receive. A logistic regression model is usually used to calculate the propensity of receiving a treatment of interest versus a comparator. For patients in the treatment group, the weight is calculated as the inverse of the propensity score, whereas for patients in the comparator group, the weight is calculated as the inverse of 1 minus the propensity score (i.e., the probability of not receiving the treatment). Thus, the weight for a given patient j with propensity score p_j is calculated as:

$$\text{weight} = \begin{cases} 1/p_j & \text{if the patient is a member of the treatment group} \\ 1/(1-p_j) & \text{if the patient is a member of the comparator group} \end{cases}$$

Once calculated, the weights determine the extent to which each patient contributes to the new pseudo-population. For example, for a patient in the treatment group with $p_j = 0.25$, the weight is $1/0.25 = 4$, which represents four units in the pseudo-population. For a patient in the comparator group with $p_j = 0.25$, the weight is $(1/(1-0.25) = 4/3)$, which represents $4/3$ units in the pseudo-population. After the pseudo-population has been created, the balance of patient characteristics should be compared across the different groups using standardized differences, as in the case for PSM. Following this, the outcomes between treatment groups are ready for comparative assessment.

Two of the case studies presented in this review implemented IPTW techniques in their CER of OACs (Table 1) [22,23]. In addition to performing PSM, Seeger and colleagues also performed IPTW analyses by re-weighting the study population of dabigatran and warfarin users using the formula given above ($1/p_j$ for dabigatran and $1/(1-p_j)$ for warfarin) [23]. Larsen and colleagues, [25] on the other hand, used generalized boosted models to re-weight the study population of apixaban, dabigatran, rivaroxaban and warfarin users. Generalized boosted models can be used to calculate weights in the case when there are three or more treatment groups being compared simultaneously [45], however, this method, and others that are based upon machine learning techniques, are beyond the scope of explanation in this introductory review [46]. A simpler alternative, when there are three or more treatment cohorts is to compute the propensity score using a multinomial logistic model with all treatment cohorts (cohort #1, cohort #2, cohort #3, cohort #4) included in the model, using one cohort as the reference (i.e., cohort #1). Each patient's weight is equal to the inverse of the probability of receiving the treatment. The weight for a given patient j with propensity score p_j is calculated as:

$$\text{weight} = 1/p_j$$

For example, for a patient with $p_j = 0.5$ in cohort #2, the weight is $1/0.5 = 2$, which represents two units in the pseudo-population. For a patient with $p_j = 0.25$ in cohort #3, the weight is $1/0.25 = 4$, which represents four units in the full pseudo-population.

A key benefit of IPTW is that all eligible patients can be analyzed. This can be particularly useful when the study population is too small to afford to lose any treated patients who could not be paired with a comparator through a matching process. From a conceptual standpoint, IPTW is somewhat more difficult to comprehend and

Table 3. A side-by-side comparison of propensity score matching and inverse probability of treatment weighting: assumptions, advantages and disadvantages.

Method	Assumptions	Pros	Cons
PSM	• No unmeasured confounding • Positivity: every subject must have nonzero probability to receive either treatment • Correct model specification for propensity score • Compare two cohorts in most cases	• Easier to understand and communicate the data	• Excludes unmatched subjects who may differ systematically from matched subjects • Limit to independent pairwise comparison
IPTW	• No unmeasured confounding • Positivity: every subject must have nonzero probability to receive either treatment • Correct model specification for propensity score • Compare two or more cohorts	• Keep all eligible subjects • Can include more than two comparisons	• Difficult to understand and communicate the data • Less intuitive • Extreme weight issue

IPTW: Inverse probability of treatment weighting; PSM: Propensity score matching.

communicate. The question of what the ATE would be in the entire population, if the treatment were offered to every member of the population, could be of interest to healthcare payers. However, oftentimes a treatment is not suitable for all members of the population. Recall the example of comparing low-dose apixaban patients with all warfarin patients, as some warfarin patients do not meet apixaban low dose criteria they are therefore not good candidates for low-dose apixaban [36]. One methodological consideration for IPTW is the issue of extreme weights. This can occur when a treated patient has an extremely low propensity score, and then a very large weight is created. Large weights can increase the variability of the estimated treatment effect, leading to potentially biased results. In order to address this, stabilized weights should be used, which trim or truncate weights to a defined threshold [47]. Seeger and colleagues encountered the issue of extreme weights, when implementing the IPTW method to compare outcomes among NVAF patients treated warfarin versus dabigatran in USA. A hazard ratio of 0.00 (0.00–0.56) for the risk of stroke and 0.08 (0.08–0.10) for the risk of major bleeding was obtained with IPTW, contrasting with PSM estimates of 0.77 (0.54–1.09) and 0.75 (0.65–0.87) respectively. The IPTW analysis was repeated after capping (truncating) the weights at a value of 4.0, resulting in IPTW estimates of 0.57 (0.46–0.71) for the risk of stroke and 0.75 (0.69–0.82) for the risk of major bleeding, which were closer (but not identical) to the PSM estimated results (0.77 [0.54–1.09] for stroke and 0.75 [0.65–0.87] for major bleeding, respectively). Lastly, whereas PSM is restricted for pair-wise comparisons of one treatment against another, IPTW is feasible to include more than two treatment groups for a comparison.

Summary

PSM and IPTW are increasingly used to address confounding by indication potentially arising in CER due to lack of randomization in treatment assignment. These methods strive to emulate the properties of an RCT in creating treatment and comparator groups with balanced distributions of patient characteristics. Both methods, while based upon the propensity score, have different interpretations, which may be more or less suitable under different scenarios. In the case where there are ample numbers of comparators available for matching to treated patients, PSM offers a more transparent method, which is readily understood and easy to communicate. Furthermore, PSM may also be more robust to misspecification of the propensity score than the IPTW method, where extreme weights can bias the estimation of the treatment effect. On the other hand, IPTW has its advantages in retaining all eligible patients in the analysis, which may be preferred if there are limitations in terms of sample size, as well as the ability to include more than two treatment comparisons simultaneously. A side-by-side comparison of the two methods is detailed in Table 3.

Conclusion

To complement RCT evidence, the comparative effectiveness and safety of the four DOACs (apixaban, dabigatran, edoxaban and rivaroxaban) versus one another and against warfarin continue to be extensively researched in real-world studies. With PSM and IPTW being increasingly used in CER of OACs, as methods to achieve balance between treatment and comparator groups, this review provides an important introduction with applied examples from the AF field to aid understanding of these methodologies. A stronger foundational basis should help researchers and end users of CER of OACs correctly interpret, communicate and implement findings into AF-related treatment decisions in clinical practice.

10.2217/cer-2020-0013

Future perspective

As CER based on real-world data gains greater prominence in healthcare decision-making, it is crucial that decision-makers are familiar with emerging methodologies used in CER to rigorously assess the strengths, limitations and quality of the evidence in order to make informed decisions.

In the field of AF research, the comparative effectiveness of available OACs continues to be extensively researched with PSM and IPTW being two increasingly applied statistical techniques. This review provides an introductory guide describing these methods side-by-side to ensure that evidence from future CER studies of OACs are critically assessed, carefully interpreted and appropriately acted upon.

Executive summary

- After decades of warfarin being the only oral anticoagulant (OAC) widely available for stroke prevention in atrial fibrillation (AF), four direct OACs (apixaban, dabigatran, edoxaban and rivaroxaban) were approved after demonstrating noninferior efficacy and safety versus warfarin in randomized controlled trials (RCTs).
- Comparative effectiveness research (CER) of OACs based upon real-world data provides complementary information to the evidence provided by RCTs.
- In real-world studies, treatments are not prescribed at random, therefore confounding by indication bias may arise in CER if there are inherent differences in the patients prescribed the two or more treatments being compared.
- Propensity score matching (PSM) and inverse probability of treatment weighting (IPTW) are increasingly popular methods used to address confounding by indication in real-world CER of OACs.
- This review was undertaken to describe the fundamentals of PSM and IPTW and presents applied examples to assist researchers and end users of CER within the atrial fibrillation field of research in critically appraising and interpreting study findings.
- Key methodological considerations -
  - PSM and IPTW methods strive to emulate the properties of an RCT in creating treatment and comparator groups with balanced distributions of patient characteristics.
  - Propensity scores form the basis for both PSM and IPTW methods, however what differs is how propensity scores are used to control for differences in characteristics of patients receiving the treatments being compared.
  - PSM works by matching patients receiving the treatment of interest with patients receiving the comparator treatment based on the estimated value of their propensity score, creating pairs of treatment and comparator patients with a similar probability of receiving treatment.
  - PSM estimates the average treatment effect for the treated, reflecting the effect of the treatment only among those who were ultimately treated.
  - In the IPTW method, weights are assigned to patients based on the inverse of their probability of receiving treatment, as estimated by the propensity score, creating a new pseudo-study population where treatment assignment is independent.
  - IPTW estimates the average treatment effect, reflecting the effect of the treatment in the scenario that every patient within the population was offered the treatment.
  - Each method has its own practical advantages as well as limitations, which may be more or less suitable under different scenarios such as the availability of treatment and comparator patients and number of treatments being compared.
  - PSM and IPTW may not always lead to the same study conclusions therefore results should be viewed with full consideration of the technical differences in methodology and differences in the measurement and interpretation of the treatment effect that each method provides.

Author contributions

All authors substantially contributed to the development and critical revision of the intellectual content and approved the final version.

Financial & competing interests disclosure

This study was funded by Bristol-Myers Squibb and Pfizer Inc. All authors are employees of the funders. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Chugh SS, Havmoeller R, Narayanan K *et al.* Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. *Circulation* 129(8), 837–847 (2014).
2. Miyasaka Y, Barnes ME, Gersh BJ *et al.* Secular trends in incidence of atrial fibrillation in Olmsted County, Minnesota, 1980 to 2000, and implications on the projections for future prevalence. *Circulation* 114(2), 119–125 (2006).
3. Ganjehei L, Massumi A, Razavi M, Rasekh A. Stroke prevention in nonvalvular atrial fibrillation. *Texas Heart Inst. J.* 38(4), 350–352 (2011).
4. Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Ann. Intern. Med.* 146(12), 857–867 (2007).
5. Granger CB, Alexander JH, Mcmurray JJ *et al.* Apixaban versus warfarin in patients with atrial fibrillation. *N. Engl. J. Med.* 365(11), 981–992 (2011).
6. Connolly SJ, Ezekowitz MD, Yusuf S *et al.* Dabigatran versus warfarin in patients with atrial fibrillation. *N. Engl. J. Med.* 361(12), 1139–1151 (2009).
7. Giugliano RP, Ruff CT, Braunwald E *et al.* Edoxaban versus warfarin in patients with atrial fibrillation. *N. Engl. J. Med.* 369(22), 2093–2104 (2013).
8. Patel MR, Mahaffey KW, Garg J *et al.* Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N. Engl. J. Med.* 365(10), 883–891 (2011).
9. Kendall JM. Designing a research project: randomised controlled trials and their principles. *Emerg. Med. J.* 20(2), 164–168 (2003).
10. Berger ML, Sox H, Willke RJ *et al.* Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol. Drug Saf.* 26(9), 1033–1039 (2017).
11. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. *J. Multidisciplin. Healthcare* 11, 295–304 (2018).
12. Li G, Lip GYH, Holbrook A *et al.* Direct comparative effectiveness and safety between non-vitamin K antagonist oral anticoagulants for stroke prevention in nonvalvular atrial fibrillation: a systematic review and meta-analysis of observational studies. *Eur. J. Epidemiol.* 34(2), 173–190 (2019).

• **Meta-analysis of existing comparative effectiveness research on oral anticoagulants.**

13. Sterne JA, Hernan MA, Reeves BC *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 355, i4919 (2016).
14. Joseph KS, Mehrabadi A, Lisonkova S. Confounding by indication and related concepts. *Curr. Epidemiol. Rep.* 1(1), 1–8 (2014).

• **Introductory reading on the concept of 'confounding by indication'.**

15. Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. *JAMA* 316(17), 1818–1819 (2016).
16. Schaumberg DA, Mcdonald L, Shah S, Stokes M, Nordstrom BL, Ramagopalan SV. Evaluation of comparative effectiveness research: a practical tool. *J. Comp. Eff. Res.* 7(5), 503–515 (2018).
17 . Borah BJ, Moriarty JP, Crown WH, Doshi JA. Applications of propensity score methods in observational comparative effectiveness and safety research: where have we come and where should we go? *J. Comp. Eff. Res.* 3(1), 63–78 (2014).

• **Commentary on the increasing application of propensity score-based methods in comparative effectivenss research.**

18. Biondi-Zoccai G, Romagnoli E, Agostoni P *et al.* Are propensity scores really superior to standard multivariable analysis? *Contemp. Clin. Trials* 32(5), 731–740 (2011).
19. Elze MC, Gregson J, Baber U *et al.* Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *J. Am. Coll. Cardiol.* 69(3), 345–357 (2017).
20. Glynn RJ, Schneeweiss S, Sturmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin. Pharmacol. Toxicol.* 98(3), 253–259 (2006).
21. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am. J. Epidemiol.* 158(3), 280–287 (2003).
22. Seeger JD, Bykov K, Bartels DB, Huybrechts K, Schneeweiss S. Propensity score weighting compared to matching in a study of dabigatran and warfarin. *Drug Saf.* 40(2), 169–181 (2017).

23. Lip GYH, Keshishian A, Li X *et al.* Effectiveness and safety of oral anticoagulants among nonvalvular atrial fibrillation patients. *Stroke* 49(12), 2933–2944 (2018).
24. Okumura Y, Yokoyama K, Matsumoto N *et al.* Three-year clinical outcomes associated with warfarin vs. direct oral anticoagulant use among Japanese patients with atrial fibrillation-findings from the SAKURA AF registry. *Circulation* 82(10), 2500–2509 (2018).
25. Larsen TB, Skjoth F, Nielsen PB, Kjaeldgaard JN, Lip GY. Comparative effectiveness and safety of non-vitamin K antagonist oral anticoagulants and warfarin in patients with atrial fibrillation: propensity weighted nationwide cohort study. *BMJ* 353, i3189 (2016).
26. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55 (1983).

•• **Propensity score original citation.**

27. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav. Res.* 46(3), 399–424 (2011).

•• **Introductory reading on propensity score matching and inverse probability of treatment weighting concepts.**

28. Gant T, Crowland K. A practical guide to getting started with propensity scores. https://support.sas.com/resources/papers/proceedings17/0689-2017.pdf
29. Leslie S, Thiebaud P. Using propensity scores to adjust for treatment selection bias. https://support.sas.com/resources/papers/proceedings/proceedings/forum2007/184-2007.pdf
30. Grotta A, Bellocco R. A review of propensity score: principles, methods, and application in Stata, Italian Stata Users' Group Meetings 2014, Stata Users Group. https://www.stata.com/meeting/italy14/abstracts/materials/it14_grotta.pdf
31. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *J. Thoracic Cardiovasc. Surg.* 134(5), 1128–1135 (2007).
32. Funk MJ, Westreich D, Wiesen C, Sturmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am. J. Epidemiol.* 173(7), 761–767 (2011).
33. Li X, Shen C. Doubly robust estimation of causal effect: upping the odds of getting the right answers. *Circ. Cardiovasc. Qual. Outcomes* 13(1), e006065 (2020).
34. Deb S, Austin PC, Tu JV *et al.* A review of propensity-score methods and their use in cardiovascular research. *Can. J. Cardiol.* 32(2), 259–265 (2016).
35. Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ* 367, l5657 (2019).
36. Nielsen PB, Skjoth F, Sogaard M, Kjaeldgaard JN, Lip GY, Larsen TB. Effectiveness and safety of reduced dose non-vitamin K antagonist oral anticoagulants and warfarin in patients with atrial fibrillation: propensity weighted nationwide cohort study. *BMJ* 356, j510 (2017).
37. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat. Med.* 33(6), 1057–1069 (2014).
38. Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *Am. J. Epidemiol.* 172(9), 1092–1097 (2010).
39. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 25(1), 1–21 (2010).
40. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceut. Stat.* 10(2), 150–161 (2011).
41. Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, Schneeweiss S. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol. Drug Saf.* 21(Suppl. 2), 69–80 (2012).
42. Bergstra SA, Sepriano A, Ramiro S, Landewe R. Three handy tips and a practical guide to improve your propensity score models. *RMD Open* 5(1), e000953 (2019).
43. Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report–Part III. *Value Health* 12(8), 1062–1073 (2009).
44. 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.* 34(28), 3661–3679 (2015).
45. Mccaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* 32(19), 3388–3414 (2013).
46. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am. J. Epidemiol.* 185(1), 65–73 (2017).
47. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health* 13(2), 273–277 (2010).