REVIEW

THE AMERICAN JOURNAL of MEDICINE®

Propensity Score Matching: The 'Devil is in the Details' Where More May Be Hidden than You Know

James A. Reiffel, MD

Columbia University, Jupiter, NY.

ABSTRACT

Propensity score matching has been used with increasing frequency in the analyses of non-prespecified subgroups of randomized clinical trials, and in retrospective analyses of clinical trial data sets, registries, observational studies, electronic medical record analyses, and more. The method attempts to adjust post hoc for recognized unbalanced factors at baseline such that the data once analyzed will hopefully approximate or indicate what a prospective randomized data set—the "gold standard" for comparing two or more therapies—would have shown. However, for practical limitations, propensity score matching cannot assess and balance all the factors that come into play in the clinical management of patients and that may be present in the circumstances of the study. Thus, propensity score matching analyses may omit, due to nonrecognition, the effects of several clinically important but not considered factors that can affect the outcomes of the analyses being reported, causing them to possibly be misleading, or hypothesis-generating at best. This review discusses this issue, using several specific examples, and is targeted at clinicians to make them aware of the limitations of such analyses when they apply their results to patients in their care.
 • The American Journal of Medicine (2019) 000:1–4

KEYWORDS: Propensity score matching; Clinical trials

A recent issue of the *European Heart Journal* contained a highly instructive paper by Davila et al[1] that compared subgroup results from the DIG trial as analyzed from randomized data with an observational non-randomized comparison that was adjusted for baseline covariates as is done for propensity score matching in many non-prospective studies. Propensity score matching attempts to adjust post hoc for recognized unbalanced factors at baseline such that the data once analyzed will hopefully approximate or indicate what a prospective randomized data set—the "gold standard" for comparing two or more therapies— would have shown. Randomization, if the patient groups are large enough, assumes that outcome-influencing factors will be equalized across study arms such that they will not have an unbalanced effect on outcome results. Propensity score matching is applied to subgroup comparisons or to "real-world" "observational" data sets, registries, analyses of electronic medical records, and the like where analyses (usually retrospective) are subject to confounding because enrollees who receive one study treatment differ systematically from those receiving another, including selection bias caused by physician choice in the treatment applied. By choosing groups based upon similar baseline demographics or other characteristics, propensity score matching attempts to mimic the effects of prospective randomization. But can it really?

A propensity has been defined as "a tendency to behave in a particular way" (Cambridge English Dictionary) or "an often intense natural inclination or preference" (Merriam Webster Dictionary). However, it is not a guarantee. It is no surprise, then, that matching baseline characteristics in an attempt to assess a propensity of the groups cannot assure that they will behave as suggested. Unfortunately, this fact is commonly overlooked by readers of trials that use propensity score matching, and the results that ensue can be misleading. The observations by Davila et al [1] that contrast

Funding: None.

Conflict of Interest: During the past 12 months JAR has served as an investigator for Janssen, an expert witness for Johnson & Johnson, and a consultant for Roivant. During the past 3 years, DAR has served as an investigator and consultant for Medtronic, Janssen, Gilead, and Sanofi; as a consultant for Portola, Acesion, and InCardia Therapeutics, and on a speaker's bureau for Janssen and Boehringer Ingelheim.

Authorship: The author is solely responsible for the content of this manuscript.

Request for reprints should be addressed to James A. Reiffel, MD, Columbia University, c/o 202 Birkdale Lane, Jupiter, FL 33458.

E-mail address: jar2@columbia.edu

0002-9343/
https://doi.org/10.1016/j.amjmed.2019.08.055

the results obtained in a post-hoc subgroup analysis with those of the full prospective randomized DIG trial are a clear example.

Their report states,[1] "The primary aim of this analysis is to assess whether adjustment in an observational analysis can lead to the same treatment effect estimate as the randomization-based analysis of the DIG trial." Unlike in the main trial where digitalis had no effect on mortality as compared with placebo, in the subgroup studied (patients who had taken digitalis prior to the trial who were then randomized to continue digitalis or switch to placebo), those previously on digitalis had a higher mortality. The authors further write, "Baseline differences between the two groups were present but adjustment for baseline population differences does not explain the observed increase in mortality." They concluded that "prescription of digoxin is an indicator of disease severity and worse prognosis, which cannot be fully accounted for by covariate adjustments in the DIG trial where patients were well-characterized. It is unlikely that weaker research approaches (observational studies of administrative data or registries) can provide more reliable estimates of the effect of cardiac glycosides." This report raises appropriate concerns regarding the differences in results obtained between prospectively collected randomized data and data obtained through propensity score matching.

CLINICAL SIGNIFICANCE

- Propensity score matching attempts to adjust post hoc for unbalanced baseline factors to mimic what a prospective randomized study would show. The method is now commonly used in retrospective study analyses.
- Propensity score matching cannot assess and balance all possible outcome-influencing factors, such as disease history and severity, drug doses, etc. Thus, propensity score matching outcomes can be misleading.
- Propensity score matching results should be interpreted with caution. They carry more bias than prospective, randomized trials in similar populations.

Propensity score matching was first described in 1983[2] and has been employed progressively since then. However, although propensity score matching has become a valuable statistical tool in clinical outcomes analysis, and is often clinically relevant, it is not all-inclusive and may not be adequately so. This is an important limitation that clinicians often do not recognize. I recently reviewed 20 randomly selected, published articles on cardiovascular health from the past 4 years that used propensity score matching.[3] In most, but not all, the factors selected for propensity score matching were listed. Some were specific to the type of trial (surgical or device, pharmacological or medical). Almost all included age, gender, weight, baseline comorbidities and concomitant disorders, classes of drugs taken by the patients, smoking, alcohol, selected blood tests and cardiac functional tests, and the like. However, virtually always, these were considered as present or absent rather than by level of severity, and, while drugs were listed by class, they were never listed by specific agent, specific dose or dose range, or drug-interaction potential. Duration of disease was also never listed. However, each of these can significantly affect outcome-trial results and hence are important limitations to propensity score matching as compared to prospective randomization. Consider the following examples discussed below:

DISEASE DURATION AND SEVERITY

The first example involves two older atrial fibrillation trials that were used to assess the efficacy and safety of sustained-release propafenone as part of its development process: RAFT (the Rythmol Atrial Fibrillation Trial, which was performed in North America) and ERAFT (the European Rythmol/Rythmonorm Atrial Fibrillation Trial, which was performed in Europe).[4,5] Both trials were prospective, randomized, double-blind, placebo-controlled trials that assessed sustained-release propafenone for atrial fibrillation. Notably, despite using the same active drug and placebo, manufactured by the same pharmaceutical company, in the same doses, with no important significant differences in baseline patient characteristics among the two arms within each trial, lower efficacy rates versus placebo were present in ERAFT than in RAFT. For example, in patients taking 425 mg twice a day, recurrence rates of atrial fibrillation at 90 days were approximately 65% in ERAFT versus approximately 30% in RAFT. Why such a difference? Importantly, patients in ERAFT had greater atrial fibrillation burden, longer atrial fibrillation history, and more prior antiarrhythmic drug failures than were present in RAFT. Equally notable, in general, both disease severity and prior antiarrhythmic drug failures predict a lower response rate to subsequent antiarrhythmic drug trials. Simply comparing ERAFT and RAFT based on the presence of enrollment-requiring atrial fibrillation and on specific concomitant diseases and comorbidities, as is common with propensity score matching, would have missed these important result-altering details. Granted, comparing two different trials is not an example of propensity score matching. However, as propensity score matching typically compares the presence and absence of listed items but rarely if ever compares them with high granularity, such as disease duration or response to prior therapies, the findings of these two trials when compared with each other show us the importance of outcome-altering differences between populations, or subgroups, that would not be accounted for in most if not all propensity score matching approaches.

SPECIFIC DRUGS WITHIN A CLASS AND SPECIFIC DRUG DOSES

Factors beyond disease severity or disease duration that may have the same type of consequences on outcomes

reported include drugs within a class and specific drug dosing. For example, while propensity score matching generally considers differences in baseline drugs regarding use or non-use, including statins, beta-blockers, angiotensin converting enzyme inhibitors, antidiabetic agents, and the like, it generally does not (if ever) consider the specific agent used within each of these classes or their doses. However, these pharmacologic specifics can have major effects on many measured clinical outcomes. For example, there are significant clinical differences among specific statins and among statin doses, and among individual beta blockers and angiotensin converting enzyme inhibitors, to name but a few commonly used drug classes in cardiovascular care.

With respect to statins, there are notable pharmacokinetic and pharmacodynamic differences among the agents.[6] These include differences in hepatic metabolism and renal excretion with consequent differences in drug interaction potential. Because many patients in cardiovascular clinical trials take multiple types of drugs on a daily basis, drug interaction potential can be important and has implications regarding efficacy and safety profiles of new agents being studied. At least equally important, multiple studies have almost consistently shown greater benefit on atherosclerotic disease consequences associated with hyperlipidemic states when high dose statins are employed versus lower dose—a consequence now recognized in clinical guidelines.[7] Such effects on outcomes can be of importance if those same outcomes are being tested in the clinical analysis to which propensity score matching is being employed but specific drugs and doses are not considered.

With respect to angiotensin converting enzyme inhibitors, again there are notable pharmacokinetic and pharmacodynamic differences and clinical outcome effects among the agents.[8] Of importance here, though rarely considered in the most recent years, there are data that strongly suggest that the agents with highest tissue penetrance, including trandolapril, ramipril, and quinapril have greater beneficial effects on endothelial function, plaque stability, and cardiovascular outcomes than those with poor tissue penetrance. Since these differences have resulted in greater benefit in, for example, post myocardial infarction studies with respect to recurrent myocardial infarction, mortality, and the like,[8] they certainly could affect the results of other trials examining similar outcomes if the baseline distribution of specific angiotensin converting enzyme inhibitors differs among groups being studied. Moreover, trandolapril has dual hepatic and renal excretion, making it safer in the presence of renal dysfunction, and, pharmacokinetic data suggest that only two truly have 24-hour duration of effect.[8]

Finally, with respect to beta-blockers, still again there are notable inter-drug differences, not only in their pharmacologic properties[9] but also in outcomes associated with their therapy.[9-11] The serum levels of hepatically metabolized beta-blockers, such as metoprolol or propranolol, can vary by up to 10-fold for the same dose; therefore a given dose in a baseline demographic profile may not mean the same clinical effect. Such variation is markedly less for renally excreted beta-blockers. Some beta-blockers have additional clinically important actions, such as alpha blockade with labetalol and carvedilol. Notably, in many direct comparative studies, carvedilol has had superior clinical outcomes compared with metoprolol, including better heart failure outcomes, better effects with respect to measurements of diabetic control, and better arrhythmia suppression.[9-16] Yet, these differences have not been considered in the analyses of studies that use propensity score matching. The same may be said about antidiabetic agents, with some classes improving outcomes in heart failure and mortality while others have not.

The above examples are but a few of many that could be chosen to suggest that propensity score matching, which has a logistical feasibility limit on the number of factors it can consider, may be clinically limited or even misleading. Thus, in my opinion, propensity score matching should be considered highly valuable but not necessarily definitive. Nor should the method be considered as reliable in balancing intergroup differences as a prospective randomization approach. In this, I am not alone. A 2008 “report card” on propensity score matching in the cardiovascular literature found “that the application of propensity-score matching in cardiology reports has been ‘poor.’”[17]

Though its results cannot be taken as proof positive with a very high degree of certainty, at a minimum they may always be viewed as hypothesis generating. Most studies employing propensity score matching acknowledge the constraints of the method via some statement in their Limitations section, such as “although adjustment was made for many variables, it is possible that residual confounders between the groups could still be present and that propensity score matching may not be able to balance all unmeasured confounders.” Such statements are statements of fact. They are appropriate to add with respect to the limitations in the analyses employed in the study and should not be overlooked. My purpose in this article is to point out to the reader that the confounders not considered in propensity score matching may have substantial clinical impact and that caution needs to be employed when considering the results of studies or when contrasting them to prospective, randomized, placebo-controlled or active-controlled trials.

REFERENCES

1. Aguirre Dávila L, Weber K, Bavendiek U, et al. Digoxin-mortality: randomized vs. observational comparison in the DIG trial. *Eur Heart J.* 2019;40(40):3336–41.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41–55.
3. Reiffel JA. Propensity-score matching: optimal, adequate, or incomplete. *J Atr Fibrillation* 2018;11(4):2130.
4. Pritchett EL, Page RL, Carlson M, Undesser K, Fava G, Rythmol Atrial Fibrillation Trial (RAFT) Investigators. Efficacy and safety of sustained-release propafenone (propafenone SR) for patients with atrial fibrillation. *Am J Cardiol.* 2003;92:941–6.
5. ERAFT Investigators, Meinertz T, Lip GY, Lombardi F, et al. Efficacy and safety of propafenone sustained release in the prophylaxis of symptomatic paroxysmal atrial fibrillation (The European Rythmol/

Rythmonorm Atrial Fibrillation (ERAFT) study. *Am J Cardiol.* 2002;90:1300–6.
6. Davidson MH, Toth PP. Comparative effects of lipid-lowering therapies. *Prog Cardiovasc Dis.* 2004;47:73–104.
7. Stone NJ, Robinson JG, Lichtenstein AH, et al. 2013 ACC/AHA Cholesterol Guideline Panel. Treatment of blood cholesterol to reduce atherosclerotic cardiovascular disease risk in adults: synopsis of the 2013 American College of Cardiology/American Heart Association cholesterol guideline. *Ann Intern Med.* 2014;160:339–43.
8. Wong J, Patel RA, Kowey PR. The clinical use of angiotensin converting enzyme inhibitors. *Prog Cardiovasc Dis.* 2004;47:116–30.
9. Reiter MJ. Cardiovascular drug class specificity: beta-blockers. *Prog Cardiovasc Dis.* 2004;47:11–33.
10. Reiffel JA. Drug and drug-device therapy in heart failure patients in the post-COMET and SCD-HeFT era. *J Cardiovasc Pharmacol Ther.* 2005;10(suppl 1):S45–58.
11. Reiffel JA. Practical algorithms for pharmacologic management of the post myocardial infarction patients. *Clin Cardiol.* 2005;28(suppl 1) [I-28-I37].
12. Bank AJ, Kelly AS, Thelen AM, Kaiser DR, Gonzalez-Campoy JM. Effects of carvedilol versus metoprolol on endothelial function and oxidative stress in patients with type 2 diabetes mellitus. *Am J Hypertens.* 2007;7:777–83.
13. Shen X, Nair CK, Aronow WS, Hee T, Esterbrooks DJ. Effect of carvedilol versus metoprolol (CR/XL on mortality in patients with heart failure treated with cardiac resynchronization therapy: a COX multivariate regression analysis. *Am J Ther.* 2013;20:247–53.
14. Merritt JC, Niebauer M, Tarakji K, Hammer D, Mills RM. Comparison of effectiveness of carvedilol versus metoprolol or atenolol for atrial fibrillation appearing after coronary artery bypass grafting or cardiac valve operation. *Am J Cardiol.* 2003;92:735–6.
15. Gilbert EM, Abraham WT, Olsen S, et al. Comparative hemodynamic, left ventricular functional, and antiadrenergic effects of chronic treatment with metoprolol versus carvedilol in the failing heart. *Circulation.* 1996;94:2817–25.
16. Ayan M, Habash F, Algam B, et al. A comparison of anti-arrhythmic efficacy of carvedilol vs metoprolol succinate in patients with implantable cardioverter-defibrillators. *Clin Cardiol.* 2019;42:299–304.
17. Austin PC. Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006. *Circ Cardiovasc Qual Outcomes.* 2008;1:62–7.