®JAFIB

Featured Review

Journal of Atrial Fibrillation

Dec 2018 - Jan 2019 Volume 11 - Issue 4

ISSN :1941-6911

JAFIB

Journal of Atrial Fibrillation

- The Biomarkers NT-proBNP and CA-125 are Elevated in Patients with Idiopathic Atrial Fibrillation.
- Current Safety of Pulmonary Vein Isolation in Paroxysmal Atrial Fibrillation First Experience of Same Day Discharge Atrial Arrhythmias in Chagas' Disease.
- Anticoagulation After Catheter Ablation of Atrial Fibrillation: Is it time to Discontinue in Select Patient Population.
- Clinical Impact of Saline Volume Infused Through Irrigated-Tip Ablation Catheter in Low Acuity Paroxysmal Atrial Fibrillation Ablation Patients.
- Gains in Paroxysmal Atrial Fibrillation Ablation Using a Standardized Workflow to Optimize Contact Force Technologies.

We Publish

• Editorials • Featured Reviews • Journal Reviews
• Original Research • Case Reports

PubMed

www.jafib.com

Published by
Cardiofront LLC

www. jafib.com

Propensity-Score Matching: Optimal, Adequate, or Incomplete?

James A. Reiffel[1]

[1]*Professor Emeritus of Medicine Columbia University.*

Abstract

The gold standard for comparing two or more therapies in modern medicine is the prospective, double-blind, randomized clinical trial. In this model, especially with sizable enrollment, randomization is assumed to equalize outcome-influencing factors across study arms such that they have no unbalanced effect on outcomes tested. Recently, clinical studies have increasingly utilized registries, electronic medical records, and other real-world observational data sets in place of or to supplement randomized trials. Notably, nonrandomized studies are subject to confounding when enrollees who receive one treatment under investigation differ systematically from those receiving another, including selection bias where patient treatments are chosen by their physicians rather than by randomization. In such studies, statistical adjustment by propensity-score matching (PSM) is commonly employed in an attempt to reduce bias from concomitant confounding variables (to correct for many baseline imbalances). PSM attempts to mimic randomization on observed covariates. PSM is also frequently used in post-hoc, retrospective, and subgroup analyses for similar reasons. Importantly, while PSM is valuable, it is not all-inclusive. It is not readily apparent that practitioners recognize this clinically important limitation and consider it when interpreting/applying study results. This paper provides supporting details regarding the above, and uses a comparison of two atrial fibrillation trials and two CHA_2DS_2-VASc circumstances to enhance the points made.

Introduction

The “gold standard” for comparing two or more therapies in modern medicine is the prospective, double-blind, randomized clinical trial. In this model, especially with sizable enrollment, randomization is assumed to equalize outcome-influencing factors across study arms such that they have no unbalanced effect on outcomes tested. Recently, clinical studies have increasingly utilized registries, electronic medical records, and other “real-world” “observational” data sets in place of or to supplement randomized trials. Notably, nonrandomized studies are subject to confounding when enrollees who receive one treatment under investigation differ systematically from those receiving another, including selection bias where patient treatments are chosen by their physicians rather than by randomization. In such studies, statistical adjustment by propensity-score matching (PSM), first described in 1983,[1] is commonly employed in an attempt to reduce bias from concomitant confounding variables (to “correct” for many baseline imbalances). PSM attempts to mimic randomization on observed covariates. PSM is also frequently used in post-hoc, retrospective, and subgroup analyses for similar reasons. Importantly, while PSM isvaluable, it is not all-inclusive. It is not readily apparent that practitioners recognize this clinically important limitation and consider it when interpreting/applying study results. Sometimes it doesn’t matter, but sometimes it can. Interestingly, a “report card on propensity-score matching in the cardiology literature” has found that the application of PSM in cardiology reports has been “poor”[2].

For this paper, I reviewed 20 randomly selected manuscripts in Circulation, JACC, and Stroke from the past 4 years. I also re-reviewed two older atrial fibrillation (AF) studies.In the 20 papers, the factors chosen for PSM were usually but not always listed. Some were specific to the type of trial, e.g., prior surgeries for surgical studies, but almost all included: age; gender; hypertension; diabetes; coronary artery disease history; heart failure or LVEF; non-ischemic heart diseases;AF +/- other rhythm detail; prior stroke; renal, hepatic, and pulmonary status; medication list; smoking; hyperlipidemia; selected blood tests, ECG findings, echocardiographic findings. Some included prior alcohol/drug abuse, weight. Notably, however, most often, these variables were considered only as present/absent and were virtually never considered in terms of severity (quantitatively). Also, rarely if ever regarded were specific drugs within a class, drug doses, drug interaction potential, or past patient history although such could significantly affect study results.

While each of the above listed comorbidities are important to recognize and adjust for by PSM with respect to their effects on study results, additional potentially confounding and results-influencing factors may be present but go unnoted/unmentioned. For example, hypertensive patients may or may not have LV hypertrophy (LVH) but the presence/absence of LVH is rarely if ever considered. Hypertensives with LVH have a poorer survival, two to four-fold greater cardiovascular (CV) morbidity, and a greater likelihood of developing AF despite antihypertensive treatment, Thus LVH may affect CV outcomes. Moreover, LVH resolution potential with antihypertensive drugs differs among the drug classes. Similarly, other

Key Words

Propensity-Score Matching, Clinical Trials, Observational Studies

Corresponding Author
James A. Reiffel,
202 Birkdale Lane Jupiter, FL 33458

factors, such as specific medications (beyond drug class), their dosing, or their possible drug interactions were only considered once in the 20 papers I reviewed. Additionally, none considered responses to prior drug trials or the duration of the comorbidities present. Consider: (1) In many trials, high dose statins have proven to be superior to lower doses in reducing major adverse cardiovascular outcomes. Yet, statin doses were not part of any propensity matching consideration that I examined. Moreover, all statins are not the same with respect to possible drug interactions. (2) Similarly, all beta blockers are not identical. Hepatically metabolized beta blockers can have up to 10-fold differences in serum concentrations and actions for a given dose, which is not the same for renally excreted beta blockers. Some have effects beyond beta receptor blockade. Some have demonstrated superiority in heart failure. Thus, simply noting beta blockers as present or absent should be clinically insufficient. (3) Likewise, it is well recognized that specific agents for diabetic management can have dramatically different effects on CV outcomes, and noting diabetes as present/absent without considering specific treatment(s) may be shortsighted. (4) The same is true for ACE inhibitors/ ARBs, where outcomes across trials have not been uniform and where tissue penetrance and effects therefrom differ among agents with differences in clinical outcomes(5) Finally disease duration and responses to prior therapy can dramatically alter treatment responses, but they are almost never considered with PSM. Here, the two older AF trials are particularly instructive. In the prospective, randomized, placebo-controlled sustained-release propafenone vs placebo AF trials, RAFT[3] and ERAFT,[4] lower efficacy rates were seen with the active drug in ERAFT vs RAFT despite using identical study drug, dose, placebo, and manufacturer for treatment of the same arrhythmia. Importantly, ERAFT had greater AF burden, longer AF history, and more prior antiarrhythmic drug (AAD) failures. Importantly, disease severity and prior AAD failure both generally predict lower response to subsequent AAD administration. Simply comparing these two populations based on the presence of prior AF and on specific underlying disease and comorbidity list would have missed these important result-altering details. Finally, consider that even the CHA2DS2-VASc score, which has been included in PSM in many trials, can be a misleading comparator for both stroke and mortality. In AF, both older age and prior stroke have a greater risk for ensuing stroke than the other CHA_2DS_2-VASc score components and age is the single strongest mortality predictor. Thus, a 69-year old female diabetic hypertensive s/p an MI likely has a lower absolute risk of both stroke and death than an 89-year old male with a prior stroke and prior MI although both patient's scores = 5. However, her risk would be higher if her hypertension had associated LVH and renal insufficiency.

In my opinion, propensity matching corrections should be considered valuable, but not clinically complete. Many papers recognize this and typically include statements in their limitations section such as: (a) "Observational studies often do not account for confounders, and the use of unadjusted values from these studies introduces bias"; (b) "Although adjustment was made for several variables, it is possible that residual confounders between the groups could have been omitted in the analysis"; (c) "Like any nonrandomized design, propensity matching may not be able to balance unmeasured confounders". Perhaps this has been best stated by Moss et al[5] : "Propensity scoring is a powerful tool that enables excellent matching of baseline characteristics, which may be superior to that obtained in a randomized trial. However, if important unobserved covariables are not identified and not entered into the propensity model, significant baseline differences may still exist between the two groups. Propensity scoring is not therefore a substitute for randomization."

Although many such confounders cannot be easily quantitated and included in a propensity score, I suggest that at least those that could be relevant to the results of the study being reported be recognized and listed, not just called residual confounders. In this way the reader will know what the investigators' PSM did not include and can reflect on their relevance, possible impacts, and the best application of the study results to his/her patients. Is it not reasonable to suggest that such an effort be made so as to further enhance the link between a clinical study and clinical practice?

Disclosures

In the past 3 years, Dr. Reiffel has been an investigator and consultant for Gilead, Janssen, and Medtronic and a consultant for Sanofi, Acesion, InCardia Therapeutics, Portola, and Roivant.

Financial support

None.

References

1. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika 1983; 70:41-55.
2. Austin PC. Report card on propensity-score matching in the cardiology literature from 2004 to 2006. Circ Cardiovasc Qual Outcomes 2008; 1:62-67.
3. Pritchett EL, Page RL, Carlson M, et al Efficacy and safety of sustained-release propafenone (propafenone SR) for patients with atrial fibrillation. Am J Cardiol 2003; 92:941-46.
4. 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 trial (ERAFT) study. Am J Cardiol 2002; 90:1300-06.
5. Moss RR, Humphries KH, Gao M, et al. Outcome of mitral valve repair or replacement: a comparison by propensity score analysis. Circulation 2003; 108(suppl II):II-90-II-97.