JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
© 2017 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER

VOL. 69, NO. 3, 2017
ISSN 0735-1097/$36.00
http://dx.doi.org/10.1016/j.jacc.2016.10.060

REVIEW TOPIC OF THE WEEK

Comparison of Propensity Score Methods and Covariate Adjustment

Evaluation in 4 Cardiovascular Studies

Markus C. Elze, PhD,[a,b] John Gregson, PhD,[a] Usman Baber, MD, PhD,[c] Elizabeth Williamson, PhD,[a] Samantha Sartori, PhD,[c] Roxana Mehran, MD, PhD,[c] Melissa Nichols, PhD,[d,e] Gregg W. Stone, MD, PhD,[d,e] Stuart J. Pocock, PhD[a]

ABSTRACT

Propensity scores (PS) are an increasingly popular method to adjust for confounding in observational studies. Propensity score methods have theoretical advantages over conventional covariate adjustment, but their relative performance in real-word scenarios is poorly characterized. We used datasets from 4 large-scale cardiovascular observational studies (PROMETHEUS, ADAPT-DES [the Assessment of Dual AntiPlatelet Therapy with Drug-Eluting Stents], THIN [The Health Improvement Network], and CHARM [Candesartan in Heart Failure-Assessment of Reduction in Mortality and Morbidity]) to compare the performance of conventional covariate adjustment with 4 common PS methods: matching, stratification, inverse probability weighting, and use of PS as a covariate. We found that stratification performed poorly with few outcome events, and inverse probability weighting gave imprecise estimates of treatment effect and undue influence to a small number of observations when substantial confounding was present. Covariate adjustment and matching performed well in all of our examples, although matching tended to give less precise estimates in some cases. PS methods are not necessarily superior to conventional covariate adjustment, and care should be taken to select the most suitable method. (J Am Coll Cardiol 2017;69:345–57) © 2017 by the American College of Cardiology Foundation.

Evaluations of therapeutic interventions generally fall into 2 categories, observational studies and randomized controlled trials (RCTs). The choice of treatment in observational studies may be influenced by patient characteristics, for example, higher-risk patients may be more or less likely to receive the intervention. Some of these differences are collected in standard databases, whereas others are not (e.g., frailty). In contrast, when studying the effect of an intervention in RCTs, confounding from both measured and unmeasured variables is avoided, and RCTs are thus generally considered the highest form of scientific investigation. Nonetheless, accurate treatment effect

From the [a]Department of Biostatistics, London School of Hygiene and Tropical Medicine, London, United Kingdom; [b]Innovative Pediatric Oncology Drug Development, F. Hoffmann-La Roche AG, Basel, Switzerland; [c]Cardiovascular Institute, Icahn School of Medicine at Mount Sinai, New York, New York; [d]Division of Cardiology, Columbia University Medical Center, New York-Presbyterian Hospital, New York, New York; and the [e]Cardiovascular Research Foundation, New York, New York. Dr. Elze is an employee of and owns stock in F. Hoffmann-La Roche AG. Dr. Mehran has received research support from Eli Lilly/Daiichi-Sankyo Inc., Bristol-Myers Squibb, AstraZeneca, The Medicines Company, OrbusNeich, Bayer, CSL Behring, Abbott Laboratories, Watermark Research Partners, Novartis Pharmaceuticals, Medtronic, AUM Cardiovascular Inc., Cardiovascular Inc., and Beth Israel Deaconess Medical Center; is a compensated member of the Janssen Pharmaceuticals and Osprey Medical Executive Committees; has financial relationships with Watermark Research Partners; is a consultant for Medscape, The Medicines Company, Boston Scientific, Merck & Company, Cardiovascular Systems Inc., Sanofi USA, Shanghai BraccoSine Pharmaceutical, and AstraZeneca; holds equity in Claret Medical and Elixir Medical; is a consultant for Medscape, The Medicines Company, Boston Scientific, Merck & Company, Cardiovascular Systems Inc., Sanofi USA, Shanghai BraccoSine Pharmaceutical, and AstraZeneca; and holds equity in Claret Medical and Elixir Medical. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

Manuscript received September 16, 2016; accepted October 19, 2016.

Listen to this manuscript's audio summary by *JACC* Editor-in-Chief Dr. Valentin Fuster.

ABBREVIATIONS AND ACRONYMS

BMI = body mass index

HPR = high platelet reactivity

IPW = inverse probability weighting

MACE = major adverse cardiovascular event(s)

PS = propensity score(s)

RCT = randomized controlled trial

estimates from observational databases can provide complementary value to RCTs. This is particularly true when RCTs enroll highly selected patients (yielding results not generalizable to all real-world scenarios), are small (because of their greater complexity and cost), or are not feasible to conduct (1).

The conventional method used to adjust for baseline differences between treatment groups in observational databases is covariate adjustment, where all relevant patient characteristics are included in a regression model relating the outcome of interest to the alternative treatments. A commonly cited concern is that such models might be overfitted when the number of covariates is large compared with the number of patients or outcome events. Although a rule of thumb is to have at least 10 events per covariate included in the model (2), more recent opinions favor relaxing this rule (3).

Propensity score (PS) methods are increasingly being used in observational studies of cardiovascular interventions as an alternative to conventional covariate adjustment; many such examples can be found published in the *Journal* (4–7). A PS is defined as the probability of a patient being assigned to an intervention, given a set of covariates (8). As the PS summarizes all patient characteristics into a single covariate, it reduces (although does not eliminate [9]) the potential for overfitting. PS methods aim to achieve some of the characteristics of RCTs by compensating for different patients having different probabilities of being assigned to the exposures under investigation. Thus, the aim of these methods is to attenuate problems of confounding of patient characteristics and assignment to an intervention typically found in observational studies.

Popular PS methods include stratification, matching, inverse probability weighting (IPW), and use of the PS as a covariate in a conventional regression model (10–12). However, there is lack of clear guidance as to how to make a sensible choice from among these various PS methods or conventional covariate adjustment for any given database. We therefore applied several PS methods to 4 large-scale observational cardiovascular datasets to critically examine the specific advantages and pitfalls of the different methods and to compare their results with those using classic covariate adjustment.

METHODS

DATASETS. We analyzed data from the CHARM (Candesartan in Heart Failure-Assessment of Reduction in Mortality and Morbidity) program (13), the ADAPT-DES study (14), the THIN study (15), and the PROMETHEUS study (16). For each dataset, we focused on 1 "treatment" comparison and 1 outcome of prime interest. The overall goal was to produce relevant PS models across a range of different settings, so for some cases these choices differed from the primary objectives of the original publications. The terms *treatment* and *control* are used throughout to simplify the language, even though 1 study (14) performed comparisons for platelet reactivity. All outcomes studied were time-to-event, with censoring occurring at the end of planned follow-up, or at the time of patient withdrawal or lost to follow-up.

The CHARM program (13) randomized 7,599 patients with chronic heart failure to candesartan or placebo therapy, with a median follow-up of 3.1 years. We investigated the association between treatment with beta-blockers at baseline (3,396 untreated, 4,203 treated) and all-cause death (1,831 events). That is, we used the CHARM program as an observational database for making inferences about the association between use of beta-blockers and risk of mortality. Our PS model contains cardiovascular risk factors (age, sex, body mass index [BMI], smoking, diabetes), as well as prior cardiovascular events and hospitalizations (18 variables in all).

The ADAPT-DES study (14) investigated the relationship between high platelet reactivity (HPR) in patients taking clopidogrel (HPR: n = 4,930; no HPR: n = 3,650) and stent thrombosis and other cardiovascular events at 12 months' follow-up in a prospective, multicenter registry of patients receiving drug-eluting stents. Herein we focused on stent thrombosis (56 events). The study authors reported an adjusted hazard ratio (HR) of 2.49 for HPR compared to patients without HPR. Our PS model contained information about age, sex, medication, diabetes, ethnicity, smoking, renal function, and other cardiovascular risk factors (39 variables in all).

The THIN population-based cohort study (15) compared 30,811 statin users with 60,921 patients not using statins, treated by the same general practitioners (total: n = 91,732) for several outcome events, including all-cause mortality (17,296 events, HR: 0.79). The inclusion criteria required at least 12 months of follow-up; thus, the first year must be excluded due to so-called immortality bias. Herein we investigated the effect of statin use on all-cause mortality. The study authors reported an adjusted HR of 0.78, comparing statin users with nonusers. Previously, a large RCT (16) in a similar patient population found an HR of 0.87. Our PS model contained cardiovascular risk factors, age, sex, BMI, smoking, drinking, other medications, and other diseases (48 variables in all).