Propensity-score matching: a clinician's guide

Andrew D. Althouse PhD

Introduction

Propensity-score matched analyses are quite common in the medical literature; for example, a quick medical literature search easily recovers several examples.[1-9] However, despite the prevalence of propensity-score matched analyses, it remains poorly understood, somewhat of a "black box" to many clinicians. This communication is intended to be an intuitive, non-technical explanation of propensity-score matching for clinical readers to (1) bolster their understanding of the procedure, allowing them to better critique studies as a reader and/or a reviewer, and (2) improve their communication with statisticians should they choose to consider propensity-score matched analyses in their own research.

Definitions of propensity score and matching

The definition of a propensity score is "the conditional probability of assignment to a particular treatment given a vector of observed covariates." [10] The same source defines matching as "…sampling from a large reservoir of potential controls to produce a control group of modest size in which the distribution of covariates is similar to the distribution in the treated group." Connecting the two statements, one may infer that propensity-score matching uses the conditional probabilities (propensity scores) to match patients from one treatment group to patients in another treatment group. The reader should note that propensity scores may also be used in other ways – such as inverse propensity score weighting – but in this communication, the focus will remain on propensity-score matched analyses.

Corresponding author: Andrew Althouse PhD
Contact Information: althousead@upmc.edu
DOI: 10.12746/swrccc2016.0415.208

Why do we use propensity-score matching?

Ideally, there would be large-scale randomized controlled trials (RCTs) available to evaluate the efficacy and effectiveness of all possible treatment strategies against all other possible treatment strategies for any given condition. However, RCTs are expensive, time-consuming, and in certain situations may be considered unethical. Furthermore, even within broad categories of a "treatment," there may be strategic decisions which do not merit an independent trial, but can be examined using data from clinical practice.

Propensity-score matched analyses attempt to replicate a randomized experiment with observational data. Randomized controlled trials (RCTs) are inherently designed to ensure that treatment groups are balanced on key risk factors; however, observational studies will often have differences between groups due to patient preferences, physician preferences, time/era effects, and other factors. An RCT eliminates the influence of patient preference, physician preference, time/era effects, and others by randomly assigning patients to a treatment approach, removing those elements from the treatment decision. However, in observational studies, it is likely that the treatment decision was driven at least in part by one or more of those factors, creating an element of confounding by indication. Consider the work of *Mohammadi et al* evaluating outcomes in bilateral internal mammary artery (BIMA) grafting for patients who underwent in-situ grafting with the radial artery

(BIMA-RA) vs. those who underwent BIMA with additional saphenous vein graft (BIMA-SVG).[6] There are likely pre-operative differences between BIMA-RA and BIMA-SVG patients which may affect the risk in each group; propensity-score matching allows the investigators to create a cohort of BIMA-RA patients and BIMA-SVG patients with comparable risk profiles.

The other valuable use of propensity-score matching is comparing groups that are not necessarily defined by treatments and/or cannot be assigned. Consider the work of Hayes et al evaluating the impact of pulmonary hypertension (PH) on patients awaiting lung transplantation.[7] Obviously PH cannot be "assigned" to patients, and most likely there will be differences between patients with and without PH; therefore, to get an accurate estimate of the risk associated with PH in this population, propensity-score matching may be used to generate a list of PH patients and non-PH patients with comparable distributions of key risk factors (i.e., age, gender, race, BMI, others relevant to a particular condition).

Outline of propensity-score matching procedure

Suppose that there are 1000 consecutive patients with a particular condition. Assume that 800 of these have received the long-established field standard (which we will refer to as "Treatment C" = control) and 200 have received a recently approved experimental approach (which we will refer to as "Treatment E" = experimental). Knowing that there are going to be pre-operative differences between the groups, we wish to create a cohort of "Treatment C" patients with similar pre-operative risk to corresponding "Treatment E" patients. The general steps to perform a propensity-score matched analysis are as follows:

1. Create logistic regression model, including all 1000 patients, with dependent variable="Received Treatment E" and potential confounders included as independent variables.
2. Compute propensity scores (conditional probability of receiving Treatment E based on covariates)
3. Match each Treatment E patient to one (or more) Treatment C patients
4. Verify that all covariates are balanced across Treatment E vs. Treatment C in the matched sample
5. Perform outcome analyses for Treatment E vs. Treatment C patients in the matched sample

Please note that even within these steps, there are a number of additional factors that may be manipulated, particularly in step #1 (selecting which variables are used in computing the propensity score) and step #3 (the matching strategy). Detailed discussion of these nuances lies beyond the scope of this article; please contact the corresponding author or consult more technically intensive references for detailed discussion.[10-15]

What propensity-score matching can do

Generally speaking, propensity-score matching can reduce two "imbalanced" groups of patients into two smaller cohorts that are approximately balanced on one or more covariates. The conceptual simplicity of having matched pairs of comparable patients from the two original groups allows researchers to appreciate the equivalence in "baseline" risk between the matched groups, and perform straightforward analyses to compare the outcomes between the matched groups. This is particularly advantageous in studies with (a) baseline imbalances between groups on many important covariates and/or (b) low number of events, in which a traditional multivariable regression model has undesirable statistical properties.

What propensity-score matching cannot do

This is arguably the most important section of this communication.

1. The test of a good propensity-score model is the degree to which it balances the measured baseline covariates between the groups.[12] However, please note that if perfect balance is not achieved, that does not mean that the statistician has erred in the matching process; more likely it means that the data may

have some of the limitations outlined below. The model must either be recreated with different parameters to achieve balance, or perhaps propensity-score methods cannot be used because of incomplete information and/or severe imbalances in the data. For detailed discussion of how to assess the match quality, please consult technical references.[13,14]

2. Propensity-score matching cannot balance groups well if there is minimal overlap between groups in one or more of the matching covariates. For example, if the "Treatment C" patients are exclusively elderly males and "Treatment E" patients are exclusively young females, one cannot create groups of "Treatment C" and "Treatment E" patients that are adequately matched just by using propensity scores. There must be at least some degree of overlap between the groups to find appropriate matches; furthermore, if there is minimal overlap between groups in the selected characteristics, propensity-score matching is not going to solve this problem.

3. There will always be some loss of patients who cannot be adequately matched. If there are 87 Treatment C patients and 82 Treatment E patients with significant differences between groups, one cannot just "do propensity score matching" and expect to get 82 Treatment C patients perfectly matched to the 82 Treatment E patients.

4. Propensity-score matching only accounts for observed covariates; factors that affect assignment to treatment and/or outcome that cannot be observed and/or measured appropriately cannot be incorporated. Consider a surgical procedure with two different access sites, Site A vs. Site B. If Site A is preferentially used for most cases while Site B is a secondary option reserved for especially difficult cases, a propensity-score matched analysis will not be able to account for that detail (unless it is somehow captured in a measurable characteristic, such as a measured size or pressure number). One may be able to match the patient on age, BMI, gender, and other measured factors, but cannot account for that unmeasurable factor. Propensity-score matching is not necessarily *useless* in this case. If one wants to perform a comparison of post-operative complications for Site B patients vs. Site A patients who are comparable on *all other* factors, that is reasonable, but one must at least remember that Site B patients had some additional complexity and (even with *all other things being equal*) would be expected to have more complications.

Conclusion

Please note that propensity scores may be used in a variety of analytic approaches, and this is far from a comprehensive guide to all that can be done using propensity scores. This communication is intended as a straightforward description of the most common application of propensity-score matching that appears in the medical literature to better inform clinicians what the procedure is, how it works, what it can do, and (perhaps most important) what it cannot do. The author sincerely hopes that this will prove useful in reading, interpreting, and analyzing the medical literature.

Author affiliation: Andrew Althouse is a biostatistician at University of Pittsburgh Medical Center, Pittsburgh, PA

Submitted: 4/1/2016
Accepted: 6/2/2016
Published electronically: 7/15/2016
Conflict of Interest Disclosures: None

References

1. Lee H, Sung K, Kim WS, Lee YT, Park SJ, Carriere KC, Park PW. Clinical and hemodynamic influences of prophylactic tricuspid annuloplasty in mechanical mitral valve replacement. J Thorac Cardiovasc Surg 2016; 151(3): 788-795.
2. Ecker BL, McMillan MT, Datta J, Mamtani R, Giantonio BJ, Dempsey DT, Fraker DL, Drebin JA, Karakousis GC, Roses RE. Efficacy of adjuvant chemotherapy for small bowel adenocarcinoma: a propensity-score matched analysis. Cancer 2016; 122(5): 693-701.
3. Thakkar B, Patel A, Mohamad B, Patel NJ, Bhatt P, Bhimani R, Patel A, Arora S, Savani C, Solanki S, Sonani R, Patel S, Patel N, Deshmukh A, Mohamad T, Grines C, Cleman M, Mangi A, Forrest J, Badheka AO. Transcatheter aortic valve replacement versus surgical aortic valve replacement in patients with cirrhosis. Catheter Cardiovasc Interv 2016; 87(5): 955-962.
4. Yang T, Lu JH, Lau WY, Zhang TY, Zhang H, Shen YN, Aishebeeb K, Wu MC, Schwartz M, Shen F. Perioperative blood transfusion does not influence recurrence-free and overall survival after curative resection for hepatocellular carcinoma: a propensity score matching analysis. J Hepatol 2016; 64(3): 583-593.
5. Park TY, Park YS. Long-term respiratory function recovery in patients with stage I lung cancer receiving video-assisted thoracic surgery versus thoracotomy. J Thorac Dis 2016; 8(1): 161-168.
6. Mohammadi S, Dagenais F, Voisine P, Dumont E, Charbonneau E, Marzouk M, Paramythiotis A, Kalavrouziotis D. Impact of the radial artery as an additional arterial conduit during in-situ bilateral internal mammary artery grafting: a propensity score-matched study. Ann Thorac Surg 2015 (Epub ahead of print)
7. Hayes Jr D, Black SM, Tobias JD, Kirkby S, Mansour HM, Whitson BA. Influence of pulmonary hypertension on patients with idiopathic pulmonary fibrosis awaiting lung transplantation. Ann Thorac Surg 2015 (Epub ahead of print)
8. Christensen TD, Skjoth F, Nielsen PB, Maegaard M, Grove EL, Larsen TB. Self-management of anticoagulant therapy in mechanical heart valve patients: a matched study. Ann Thorac Surg 2015 (Epub ahead of print)
9. Lee PC, Kamel M, Nasar A, Ghaly G, Port JL, Paul S, Stiles BM, Andrews WG, Altorki NK. Lobectomy for non-small cell lung cancer by video-assisted thoracic surgery: effects of cumulative institutional experience on adequacy of lymphadenectomy. Ann Thorac Surg 2015 (Epub ahead of print)
10. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41-55.
11. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics 1996; 52: 249-264.
12. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statistics in Medicine 2008; 27: 2037-2049.
13. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine 2009; 28: 3083-3107.
14. Belitser SV, Martens EP, Pestman WR, Groenwold RHH, de Boer A, Klungel OH. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf 2011; 20: 1130-1137.
15. Ali MS, Groenwold RHH, Belitser SV, Pestman WR, Hoes AW, Roes KCB, de Boer A, Klunger OH. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol 2015; 68: 122-131.