Statistics in Brief: An Introduction to the Use of Propensity Scores

Maria C. S. Inacio PhD, Yuexin Chen BS, Elizabeth W. Paxton MA,
Robert S. Namba MD, Steven M. Kurtz PhD, Guy Cafri PhD, MStat

Received: 23 September 2014 / Accepted: 27 February 2015 / Published online: 13 March 2015
© The Association of Bone and Joint Surgeons® 2015

Background

Randomized controlled trials (RCTs) are considered the gold standard of clinical research because randomization reduces the risk of extraneous factors influencing results of a study [2]. Nonetheless, high-quality, observational studies are at times more desirable than experimental studies (such as RCTs) owing to their capacity to evaluate rare events, fewer ethical challenges with conducting the study, feasibility attributable to lower costs or infrastructure needs, and sometimes greater generalizability of the findings because of less-strict inclusion or exclusion criteria for patients and surgeons. Most orthopaedic studies are observational and retrospective [4, 7, 24].

Each author certifies that he or she, or a member of his or her immediate family, has no funding or commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.

This study was performed at Kaiser Permanente, San Diego, CA, USA.

M. C. S. Inacio (✉), Y. Chen, E. W. Paxton, G. Cafri
Surgical Outcomes and Analysis, SCPMG Clinical Analysis,
Kaiser Permanente, 8954 Rio San Diego Drive, Suite 406,
San Diego, CA 92108, USA
e-mail: maria.cs.inacio@kp.org

R. S. Namba
Department of Orthopedics, Orange County Kaiser Permanente,
Irvine, CA, USA

S. M. Kurtz
Exponent Inc, Drexel University, Philadelphia, PA, USA

Confounding exists when a third variable, which is not the exposure or outcome of interest, changes the relationship between the exposure and outcome being studied. For a variable to be a confounder, it must be (1) associated with the exposure of interest in the study, and (2) associated with the outcome. For a more real-life example, consider a study evaluating differences in time to revision between ceramic-on-ceramic and metal-on-polyethylene bearings used in THAs. The surgeon's choice of bearing surface is likely not random; younger patients preferentially receive ceramic-on-ceramic bearings as opposed to metal-on-polyethylene bearings. If age also is related to the outcome (revision) in the study population, then age is regarded as a confounder. If the effect of age is not incorporated in the analysis, the estimate of the treatment effect (eg, odds ratio) will be biased. Confounding can be addressed using several methods with similar objectives during either the design or analysis phases of a study. Examples of methods used during the design of a study include restriction or matching, whereas those used during analysis include stratification, regression adjustment, instrumental variables techniques, and propensity score techniques.

The purpose of this article is to describe confounding and how its effects can be minimized in observational studies with propensity score techniques. We provide guidance for when and how to use propensity scoring in studies.

What are Propensity Scores and When Should They Be Used in Orthopaedic Research?

Propensity scores are an alternative method to estimate the effect of receiving treatment when random assignment of treatments to subjects is not possible. They should be used

in orthopaedics when it is not feasible to randomize patients to different treatments. Briefly, the propensity score is the conditional probability of someone having a specific treatment given a set of variables known about this person [3]. One example might be the probability that a person receives a certain implant based on their age, sex, and indication for surgery. The fundamental idea behind propensity score methods is that cases with the same propensity score will be comparable with respect to covariates used to calculate the score, so it is only a matter of chance between treatments [3, 5, 6, 29]. When cases are comparable with respect to covariates, the effect reproduces that induced by randomization in a clinical trial. A good take-home is that propensity score techniques can allow one to mimic some of the characteristics of an RCT in the context of an observational study.

Different propensity score techniques use these conditional probabilities in different ways. For example, similar to the idea of matching patients on specific characteristics, propensity score techniques can match patients on their likelihood of being in a certain group, but without being limited to just a few variables. Propensity score techniques are not necessary in all studies. For example, studies of TKA component features (eg, rotation or bearing types) have been done where one patient gets two different components (ie, one in each knee). In this setting, variables such as age, sex, or BMI that could affect the relationship between exposure (whatever design feature you were trying to compare in the specific knee construct) and outcome are the same between the two groups. Here, there is no need to use propensity score techniques. However, when a large number of confounders are present and/or the number of events (outcomes) is small, the use of the propensity score would be preferable.

How are Propensity Scores Used?

Step 1. Calculating the Propensity Score

The propensity score is the probability of receiving one of the treatments being compared, given the measured covariates. Covariates are the variables included in the study that are not the outcome or the exposure of interest; they could be confounders or not. The propensity score is calculated by fitting a logistic regression model with treatment received as the dependent variable. A logistic regression model measures the change in likelihood of a specific dependent variable given a set of independent variables. For example, supposed we are interested in estimating the probability of someone getting a unicompartmental knee arthroplasty compared to a total knee arthroplasty. The outcome here is the actual treatment they are having and

the predictor variables of interest are their age, activity level, and osteoarthritis severity. This technique can be performed using any currently available statistical software package. The estimated propensity score provides one score for each research subject and summarizes the information about all the variables of interest.

Step 2. Checking for Propensity Score Balance

Typically, a standardized difference for each covariate is calculated before and after applying the propensity score adjustment. Rubin, a pioneer in field of propensity scores, set forth guidelines for global assessment of balance between covariates [22]. Balance assessment should correspond to how the data ultimately are analyzed.

Step 3. Using the Propensity Score in the Analysis

The propensity score can be used in several different ways for analysis. Fundamentally, the investigator needs to decide whether to assess the average treatment effect or the average treatment effect on the treated. Although these appear similar, they are distinct entities. In a comparison of two treatments, the average treatment effect is the average effect on all individuals (ie, the effect of moving all individuals from one treatment to another), this means all patients hypothetically could be candidates for either treatment. The average treatment effect on the treated is the average effect of treatment on the individuals who receive only the treatment of interest [11]. The average treatment effect on the treated applies to cases where treatment is a more narrowly targeted treatment, or one that may be difficult to adopt by all patients. An example of when average treatment effect on the treated would be estimated is when studying the effect of unicompartmental knee arthroplasty devices on the risk of revision surgery. Estimating average treatment effect on the treated is the most appropriate for addressing this example because the use of unicompartmental knee arthroplasty devices (our treatment) is applicable to a restricted group of individuals. Specifically, the unicompartmental knee arthroplasty is indicated only for individuals who have localized and single-compartment osteoarthritis. It would have been preferable to compare only individuals with the indication of single-compartment osteoarthritis for treatment; however, this level of detail typically is absent from some datasets. Another motivation for average treatment effect on the treated estimation is that unicompartmental knee arthroplasty devices are not a widely available treatment option and therefore may not be relevant for all patients who would be candidates for knee arthroplasty (unicompartmental or otherwise). Conversely, if either treatment is equally easy to adopt by

patients (or implement by surgeons), then average treatment effect would be most appropriate. An example of when average treatment effect would be estimated would be when studying the effect of highly crosslinked polyethylene inserts compared with conventional polyethylene inserts in the risk of revision knee arthroplasty. Both of these inserts technically can be used in all patients who are candidates for knee arthroplasty and therefore an average treatment effect can be estimated.

After deciding whether average treatment effect or average treatment effect on the treated should be used, one or more of the following approaches can be taken to implement the propensity score:

- (1) **Matching:** This estimates average treatment effect on the treated only. The most common implementation of propensity score matching is one-to-one or pair matching, in which pairs of treated and untreated subjects are formed, such that matched subjects have similar values of the propensity score. Although one-to-one matching appears to be the most common approach to propensity score matching, other approaches can be used. Variations include generating multiple matches and matching with replacement [28].
- (2) **Stratification (or subclassification):** This estimates either average treatment effect or average treatment effect on the treated. Stratification on the propensity score involves stratifying subjects into mutually exclusive subsets based on their estimated propensity score. A common approach is to divide subjects into five equal-size groups or strata using the quintiles of the estimated propensity score. The study by Pugely et al. [18] of the effect of general versus spinal anesthesia on short-term complication risk after primary total knee arthroplasty is an example of using stratification on propensity scores quintiles to adjust their estimates for the effect of group imbalances. A treatment effect can be estimated in each strata (cases with similar propensity score will have comparable covariate profiles), then combined across strata using weights to obtain an overall estimate. If the weights are based on equal weighting of the equally sized strata, then this estimates average treatment effect; whereas if it is based on the proportion treated in each stratum, this estimates the average treatment effect on the treated [11].
- (3) **Weighting:** This estimates either average treatment effect or average treatment effect on the treated. In propensity score weighting, the treated and control observations are reweighted to make them more representative of the population. With average treatment effect on the treated weighting, individuals in

the treated condition of interest are given a weight of 1 and individuals in the other treatment are given weights based on the odds of the propensity score to weigh up to the treatment group of interest. More often, however, weights are applied to estimate average treatment effect. In the weighted average treatment effect approach, each subject receives a weight that is the inverse probability of being in the group they are in. Weights restore balance in the distribution of the measured covariates that one would have achieved if subjects originally were randomized into treatments, which is a similar concept to that of using weighting in survey sampling. As with stratification, no restriction is placed to ensure common support, which refers to the degree of overlap in the propensity score distributions of the treatment groups. However, one method, marginal mean weighting through stratification, does include such restrictions [9, 10]. Marginal mean weighting through stratification also should be considered because it is less likely subject to misspecification of the functional form of the propensity score model than inverse probability of treatment weighted [9].

- (4) **Regression Adjustment:** This estimates the average treatment effect. This approach involves including the propensity score in the model as a covariate. It is not advocated because it requires correct specification of the functional form of the propensity score. However, it is used at times in combination with one of the previously described approaches (matching, stratification, or weighting) to remove any residual differences between treatment groups [23]. By convention, consider adding covariates to the model when residual standardized differences are greater than 0.1 [3].

Before choosing the propensity score analytic approach, one should consider whether average treatment effect or average treatment effect on the treated is more relevant and what is feasible given the data. Choose the approach that provides maximum balance on the covariates and thereby minimizes bias. If balance is comparable in several approaches, choose one that minimizes the variance of the treatment effect estimate.

Question: What Role can Propensity Scores Play in Orthopaedic Research?

Propensity score techniques were introduced in 1983 [21], and to the best of our knowledge, first used in an article in an orthopaedic journal in 2008 [16]. In a brief review of the contemporary orthopaedic literature (2000 to July 2013,

PubMed, English-language only), we found more than 30 articles using the propensity score in their analyses. Thirteen [1, 8, 12–15, 17–20, 25–27] of the 30 studies were published during the past 3 years, suggesting a growing trend.

Propensity score techniques may be the closest method we have to quasirandomization [5] in observational research. With the increased sophistication of observational cohort studies being conducted, accompanying sophistication of the analytic tools used to evaluate collected data is required. Although propensity scores are not appropriate for every analysis, the techniques offer benefits with specific conditions and should be considered when choosing analytical techniques. Propensity score techniques provide investigators and readers with increased assurance that study conclusions are the result of differences in treatment rather than differences in study groups.

Myths and Misperceptions

- Using the propensity score will remove all confounding and therefore will allow causal conclusions in your analysis.

In fact, the propensity score can adjust only for measured confounders, which are variables for which you have information and have included in your analysis.

- Using the propensity score guarantees measured confounding has been removed.

Residual confounding is possible and should be investigated. One should review and evaluate all variables included in the propensity score calculation to assure that known confounders (ie, important variables according to the literature) were not omitted.

- The objective of the propensity score model is to predict treatment as well as possible.

Models that calculate the propensity score may not be great predictive models (eg, have high concordance indices). The objective of the propensity score model is to produce a propensity score that will create the most balance between treatment groups for each confounder.

- All available predictors of treatment should be included in the propensity score model.

Only potential confounders should be included, and not variables related to only the treatment.

- The propensity score should be used whenever possible.

There are advantages of using the propensity score (such as studying rare events and accounting for a large number of confounders), but use of the propensity score involves assumptions, including that the propensity score model is correctly specified. Disadvantages include the need for large sample sizes and for

substantial overlap between groups. Finally, propensity scoring cannot account for or uncover unobserved variables.

Conclusion

Propensity scoring is a statistical method which allows investigators to mimic some of the characteristics of an RCT in the context of an observational study. Limitations include the need for a large sample size, need for substantial overlap in terms of variables for the groups under consideration, and the lack of a gold standard regarding which characteristics should be included in its estimation. However, when performed properly, propensity scoring is a useful tool which provides increased likelihood that the effects that you see in an observational study are causal.

Conflict of interest Dr. Kurtz reports that he is an employee and shareholder of Exponent, Inc., and that institutional support is received as a PI from Smith & Nephew; Stryker; Zimmer; Biomet; Depuy Synthes; Medtronic; InVivo; Stelkast; Formae; Kyocera Medical; Wright Medical Technology; Ceramtec; DJO; Celanese; Aesculap; Spinal Motion; and Active Implants, outside the submitted work.

References

1. Aaltonen KJ, Virkki LM, Jansen E, Sokka T, Kontinen YT, Peltomaa R, Tuompo R, Yli-Kerttula T, Kortelainen S, Ahokas-Tuohinto P, Blom M, Nordstrom DC. Do biologic drugs affect the need for and outcome of joint replacements in patients with rheumatoid arthritis? A register-based study. *Semin Arthritis Rheum*. 2013;43:55–62.
2. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schünemann HJ, Edejer T, Varonen H, Vist GE, Williams JW Jr, Zaza S; GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ*. 2004; 328:1490.
3. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011; 46:399–424.
4. Cunningham BP, Harmsen S, Kweon C, Patterson J, Waldrop R, McLaren A, McLemore R. Have levels of evidence improved the quality of orthopaedic research? *Clin Orthop Relat Res*. 2013;471:3679–3686.
5. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998; 17:2265–2281.
6. Faries DE, Leon AC, Haro JM, Obenchain RL. *Analysis of Observational Health Care Data Using SAS*. Cary, NC: SAS Institute; 2010.
7. Hanzlik S, Mahabir RC, Baynosa RC, Khiabani KT. Levels of evidence in research published in The Journal of Bone and Joint Surgery (American Volume) over the last thirty years. *J Bone Joint Surg Am*. 2009; 91:425–428.

8. Hellsten EK, Hanbidge MA, Manos AN, Lewis SJ, Massicotte EM, Fehlings MG, Coyte PC, Rampersaud YR. An economic evaluation of perioperative adverse events associated with spinal surgery. *Spine J*. 2013; 13:44–53.
9. Hong G. Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *J Educ Behav Stat*. 2010; 35:499–531.
10. Hong G. Marginal mean weighting through stratification: a generalized method for evaluating multivalued and multiple treatments with nonexperimental data. *Psychol Methods*. 2012; 17:44–60.
11. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004; 86:4–29.
12. Inacio MC, Cafri G, Paxton EW, Kurtz SM, Namba RS. Alternative bearings in total knee arthroplasty: risk of early revision compared to traditional bearings: an analysis of 62,177 primary cases. *Acta Orthop*. 2013; 84:145–152.
13. Lovald ST, Ong KL, Lau EC, Schmier JK, Bozic KJ, Kurtz SM. Mortality, cost, and downstream disease of total hip arthroplasty patients in the medicare population. *J Arthroplasty*. 2014;29:242–246.
14. Mandel S, Schilling J, Peterson E, Rao DS, Sanders W. A retrospective analysis of vertebral body fractures following epidural steroid injections. *J Bone Joint Surg Am*. 2013;95:961–964.
15. Margolis DJ, Gupta J, Hoffstad O, Papdopoulos M, Glick HA, Thom SR, Mitra N. Lack of effectiveness of hyperbaric oxygen therapy for the treatment of diabetic foot ulcer and the prevention of amputation: a cohort study. *Diabetes Care*. 2013; 36:1961–1966.
16. Paus AC, Steen H, Roislien J, Mowinckel P, Teigland J. High mortality rate in rheumatoid arthritis with subluxation of the cervical spine: a cohort study of operated and nonoperated patients. *Spine (Phila Pa 1976)*. 2008; 33:2278–2283.
17. Petilon JM, Glassman SD, Dimar JR, Carreon LY. Clinical outcomes after lumbar fusion complicated by deep wound infection: a case-control study. *Spine (Phila Pa 1976)*. 2012; 37:1370–1374.
18. Pugely AJ, Martin CT, Gao Y, Mendoza-Lattes S, Callaghan JJ. Differences in short-term complications between spinal and general anesthesia for primary total knee arthroplasty. *J Bone Joint Surg Am*. 2013; 95:193–199.
19. Pugely AJ, Martin CT, Gao Y, Mendoza-Lattes SA. Outpatient surgery reduces short-term complications in lumbar discectomy: an analysis of 4310 patients from the ACS-NSQIP database. *Spine (Phila Pa 1976)*. 2013; 38:264–271.
20. Raphael IJ, Tischler EH, Huang R, Rothman RH, Hozack WJ, Parvizi J. Aspirin: an alternative for pulmonary embolism prophylaxis after arthroplasty? *Clin Orthop Relat Res*. 2014;472: 482–488.
21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
22. Rubin DB. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv Outcomes Res Methodol*. 2001; 2:169–188.
23. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc*. 2000;95:573–585.
24. Samuelsson K, Desai N, McNair E, van Eck CF, Petzold M, Fu FH, Bhandari M, Karlsson J. Level of evidence in anterior cruciate ligament reconstruction research: a systematic review. *Am J Sports Med*. 2012; 41:924–934.
25. Seicean A, Seicean S, Alan N, Schiltz NK, Rosenbaum BP, Jones PK, Kattan MW, Neuhauser D, Weil RJ. Preoperative anemia and perioperative outcomes in patients who undergo elective spine surgery. *Spine (Phila Pa 1976)*. 2013; 38:1331–1341.
26. Seicean A, Seicean S, Alan N, Schiltz NK, Rosenbaum BP, Jones PK, Neuhauser D, Kattan MW, Weil RJ. Effect of smoking on the perioperative outcomes of patients who undergo elective spine surgery. *Spine (Phila Pa 1976)*. 2013; 38:1294–1302.
27. Shi HY, Chang JK, Chiu HC. Volume associations in total hip arthroplasty: a nationwide Taiwan population-based study. *J Arthroplasty*. 2013;28:1834–1838.
28. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010; 25:1–21.
29. Stuart EA, Marcus SM, Horvitz-Lennon MV, Gibbons RD, Normand SL. Using non-experimental data to estimate treatment effects. *Psychiatr Ann*. 2009; 39:41451.