Propensity scores in surgery: don't believe the hype

Richard Stevens, Jason Oke.

Observational studies and confounding

Propensity scores have become increasingly popular in surgical journals. A google scholar search using terms "Propensity score" AND "surgery" yields in excess of 17,000 results since 2021 alone (1). What is propensity score matching and is the enthusiasm for this technique in the realm of surgery-based research justified?

The propensity score is a statistical technique that is used to reduce biases inherent in observational research. Without random assignment, two groups of patients defined by some surgical interventions are likely to be different in a number of important variables. A direct comparison would not be comparing like with like. For example, in an observational study comparing post-operative outcomes in patients undergoing laparoscopic repair versus open repair, age may be associated with the type of surgery, and it may also affect the likelihood of a good post-operative outcome. The average age of patients undergoing open repair will differ from those undergoing laparoscopic repair and as a consequence, the incidence of post-operative outcomes will be artificially lower in one group compared to the other. Age is therefore a "confounder": a common cause of both the exposure and the outcome. To make an unbiased comparison of the two types of surgery, it is necessary to account for age, and any other differences that may "confound" the comparison.

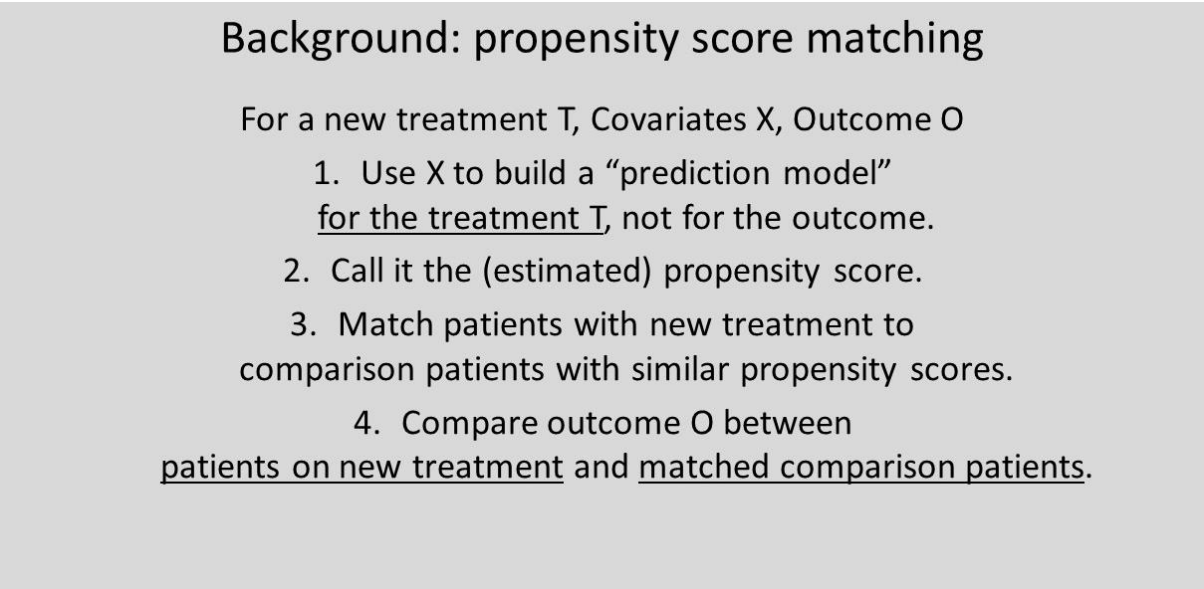
The problem of confounding can be addressed in more than one way. Arguably, the simplest way is to match patients on the observed variables. To match by age, we would for each patient undergoing open repair, find a patient of the same age who had undergone laparoscopic repair. This allows a comparison that is unconfounded by age differences. Stratification is a related method that broadly matches whole groups. Methods also exist to match on multiple variables at once. Alternatively, a regression analysis can control for confounding using mathematical calculations. An important difference between regression and matching is that regression uses the whole dataset – for better or for worse – whereas matching excludes any patients that cannot be matched, to similar patients in the other group, are excluded (2). The pros and cons of excluding much of the data, when matching, have been debated elsewhere (3).

What are propensity scores

One way to match on multiple variables, when there is more than potential confounder, is called "propensity score matching". It proceeds as follows. First, build a regression model for the relationship between the confounders and the treatment: for example, an equation for the relationship of age, and other relevant factors, to the decision to perform laparoscopic vs. open surgery. Second, for every patient in the data set, use their data and the model from the first step to estimate their "propensity" for treatment. The number assigned to each patient is usually referred to as the propensity score, although strictly speaking it is an estimated propensity score. In the third step, a matched data set is created: each patient who received treatment is matched to an untreated treatment with a similar propensity score. Finally, the outcomes of surgery in the treatment group and the matched comparison group are compared. Notice that the outcome we wish to study is not used in the analysis until the final step (see Box).

We think this method of analysis, "propensity score matching", is not very different from other approaches to matching (4). It aims to overcome confounding by making comparisons between

groups who differ in their treatment but are similar, or “balanced”, in other ways (5). As with any method of creating a matched study, this may reduce the size of the data set: matches are not available for all patients.



Other propensity score methods exist including adjustment, weighting and stratification; all follow the first two steps described above, but use different methods in the final stages (5).

Why are they useful in surgery

It can be challenging to do randomised control trials of surgery. It can be unethical and often impractical to randomise patients to different surgical interventions and urgent or life-threatening conditions present challenges with consent and randomisation in trials (6). Whilst blinding of participants and outcome assessors in surgical trials can often be achieved, blinding of the surgeon for evaluations of surgical techniques is much harder (7). Finally, there may be less commercial incentive for a company to fund randomised surgical trials (8) . As a consequence, interventions in general surgery are half as likely to be based on RCT evidence as treatments in internal medicine (6) .

Don’t believe the hype

Propensity score methods have been said to "mimic” randomized controlled trials (9, 10) . Such claims are overstated. The power of randomised trials is that over all possible randomisations, they balance all confounders, even confounders that are not measured (11). A propensity score matched study, like other matched or adjusted observational studies, can only address the confounders that are known and have been measured; it is no more a “randomized trial” than any other matched, or adjusted, observational study (12).

We advise journal readers to examine critically other claimed advantages of propensity scores over other methods of matching and adjustment. The ability to estimate the marginal, rather than conditional, odds ratio is often claimed as an advantage of propensity score matching over regression adjustment (13). This issue consumes pages of statistical journals but we think it is uninteresting to clinical readers. Would you discount a study of a new treatment because it reported odds ratios, instead of risk ratios? If not, the distinction between studies using different types of odds ratio should concern you still less.

Certainly, propensity score matching (and propensity score methods in general) can cope better than multivariate adjustment when the outcome of interest is particularly rare (14); but other matching strategies can also address this challenge (15).

The theoretical advantages of propensity scores cited by their enthusiasts, as discussed above, make little difference in practice. A systematic review of 43 studies and 79 exposure outcome associations compared the odds or hazard ratios derived using both methods and found that statistical significance differed in only 8 of the associations (16). In all 8 cases, the propensity score gave non-significant findings but traditional methods gave a statistically significant finding. This is perhaps consistent with King's argument that propensity score matching is inefficient (15).

Drawbacks of propensity scores

Propensity score matching is inconvenient (not impossible) to combine with multiple imputation, a popular method for addressing missing data. We think propensity scores also have other drawbacks. One of the attractions of matching, rather than regression adjustment or weighting, is transparency: it is easy to interpret a comparison between two groups who have been matched on age, sex, weight and diabetes status. We think this advantage is lost when the groups have instead been matched on an "estimated propensity score", where two people of quite different ages may be matched because, say, the older age of one is offset by the higher BMI of the other. Mathematical proofs exist that this kind of matching can give valid results, but the appealing transparency of matching has been lost.

What to look for in a paper using prop scores

Checklists exist for practice and reporting of propensity score studies (17). We particularly recommend looking for details of how matching was applied, how it affected sample size, and what checks were carried out that the matched populations were similar. Look for details of the factors included in the propensity score, being sure that it did not include covariates measured later in time and/or covariates that are consequences, rather than indicators, of treatment. Such errors can cause surprisingly large biases(18).

Recommendation

There is not room in an editorial to address every use of propensity scores, and we have restricted attention to propensity score matching. With that limitation, we recommend that users of the surgical literature view propensity score matching as little different to other forms of matching.

Invitation

In our search for a problem in medical science that can be usefully solved by propensity scores, but not by other methods, we have made a game of the challenge. In 2016 a modest prize was placed in a locked box, held by one of the authors, while the key is held by the other. The prize will be unlocked when someone brings a case study that convinces our statistics team, in a vote. Entrants should be careful that their proposed case study is not a comparison of one propensity score method to another; nor a comparison restricted to limited competing methods (e.g. propensity score matching compared to multivariate adjustment); nor an argument based on marginal vs. conditional odds ratios, unless you can first persuade us that the distinction matters. Author JO looks forward to receiving your submissions.

References

1. Search for “Propensity score” AND “surgery” Google Scholar; 2022 [cited 2022 09/02/2022]. Available from: https://scholar.google.com/scholar?as_ylo=2021&q=%E2%80%9CPropensity+score%E2%80%9D+AND+%E2%80%9Csurgery%E2%80%9D+&hl=en&as_sdt=0,5.
2. Brazauskas R, Logan BR. Observational Studies: Matching or Regression? Biol Blood Marrow Transplant. 2016;22(3):557-63.
3. Rosenbaum PR, Rubin DB. The Bias Due to Incomplete Matching. Biometrics. 1985;41(1):103-16.
4. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press; 2006.
5. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011;46(3):399-424.
6. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. BMJ. 2002;324(7351):1448-51.
7. Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. Trials. 2009;10:9.
8. Wente MN, Seiler CM, Uhl W, Buchler MW. Perspectives of evidence-based surgery. Dig Surg. 2003;20(4):263-9.
9. Ross ME, Kreider AR, Huang YS, Matone M, Rubin DM, Localio AR. Propensity Score Methods for Analyzing Observational Data Like Randomized Experiments: Challenges and Solutions for Rare Outcomes and Exposures. American Journal of Epidemiology. 2015;181(12):989-95.
10. Moons P. Propensity weighting: how to minimise comparative bias in non-randomised studies? European Journal of Cardiovascular Nursing. 2020;19(1):83-8.
11. Senn S. Testing for baseline balance in clinical trials. Statistics in Medicine. 1994;13(17):1715-26.
12. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. Ann Intern Med. 2002;137(8):693-5.
13. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. International Journal of Epidemiology. 2008;37(5):1142-7.
14. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol. 2003;158(3):280-7.
15. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. Political Analysis. 2019;27(4):435-54.
16. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. Journal of Clinical Epidemiology. 2005;58(6):550-9.
17. Yao XI, Wang X, Speicher PJ, Hwang ES, Cheng P, Harpole DH, et al. Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies. Journal of the National Cancer Institute. 2017;109(8).
18. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: Example using statins for preventing progression of diabetes. BMJ (Online). 2010;340(7752):907-11.