Paper 3160-2016

Using Administrative Databases for Research: Propensity Score Methods to Adjust for Bias

Ruth Croxford, Institute for Clinical Evaluative Sciences

ABSTRACT

Health care and other programs collect large amounts of information in order to administer the program. A health insurance plan, for example, probably records every physician service, hospital and emergency department visit, and prescription medication—information that is collected in order to make payments to the various providers (hospitals, physicians, pharmacists). Although the data are collected for administrative purposes, these databases can also be used to address research questions, including questions that are unethical or too expensive to answer using randomized experiments. However, when subjects are not randomly assigned to treatment groups, we worry about assignment bias—the possibility that the people in one treatment group were healthier, smarter, more compliant, etc., than those in the other group, biasing the comparison of the two treatments. Propensity score methods are one way to adjust the comparisons, enabling research using administrative data to mimic research using randomized controlled trials. In this presentation, I explain what the propensity score is, how it is used to compare two treatments, and how to implement a propensity score analysis using SAS®. Two methods using propensity scores are presented: matching and inverse probability weighting. Examples are drawn from health services research using the administrative databases of the Ontario Health Insurance Plan, the single payer for medically necessary care for the 13 million residents of Ontario, Canada. However, propensity score methods are not limited to health care. They can be used to examine the impact of any nonrandomized program, as long as there is enough information to calculate a credible propensity score.

INTRODUCTION

While randomized controlled trials (RCTs) are considered to be the “gold standard” when it comes to providing evidence of treatment efficacy, they often have drawbacks. Due to strict inclusion criteria, they may lack generalizability. The high cost of conducting an RCT affects the sample size and follow-up period, which has an impact on their ability to detect relatively rare but serious outcomes. And there are many interesting research questions which are impractical or unethical to study using a randomized design.

Observational data, which include data collecting using chart reviews and administrative data, can be used to compare treatments relatively quickly and inexpensively, and can be used to research questions which cannot be addressed using randomized trials. In the case of research using administrative databases, the results can be generalized to the full range of patients and providers covered by the data. However, in studies using observational data, the researcher has no control over the treatment assignment. This gives rise to concerns about treatment bias – the possibility that one of the treatments was given to people who were, on average, sicker or harder to treat. Propensity score methods are one way to adjust, analytically, for possible differences between two treatment groups, allowing inferences to be made about treatment effectiveness.

This paper describes how the propensity score is calculated, and how it is used to compare two treatment groups using either a sample matched on the propensity score, or a sample weighted by the inverse probability of treatment. Calculation of the propensity score and creation of a matched sample are illustrated using a study conducted to compare two drugs. While the example is drawn from health services research, propensity score methods are useful for anyone who would like to compare two groups in the absence of random assignment.

RANDOMIZED CONTROLLED TRIALS

An RCT is a comparison of two treatments in which treatment assignment is determined by a randomization process. As well, researchers, raters, and subjects are usually blinded to the treatment assignment until the end of the trial.

Figure 1 illustrates the basic elements of an RCT. It begins with the population of interest, shown in the diagram as a collection of healthier (smiling faces) and less healthy (frowning faces) individuals. Exclusion and inclusion criteria are used to restrict the study population. In the diagram, individuals selected for the study are shown as including only the healthier members of the population, a point that I will come back to later. The study has a well-defined baseline time point – the time at which the treatment choice is made. For the sake of illustration, the choice is shown as being made by a coin toss, but in reality, selection is made using reproducible random allocation. The two treatment groups are then followed up for a pre-specified time. At the end of the follow-up period, each subject's outcome is determined.

RCTs have a number of important characteristics, which should be kept in mind when the propensity score methods are described. The inclusion and exclusion criteria ensure that all of the subjects are candidates for both treatments. The randomization process ensures that the two treatment groups are comparable in all respects. Not only can it be demonstrated that there are no important differences between the two groups with respect to all of the measured covariates, but the two groups are also similar with respect to unmeasured covariates. Thus, any difference in outcomes can be attributed to differences in treatment.

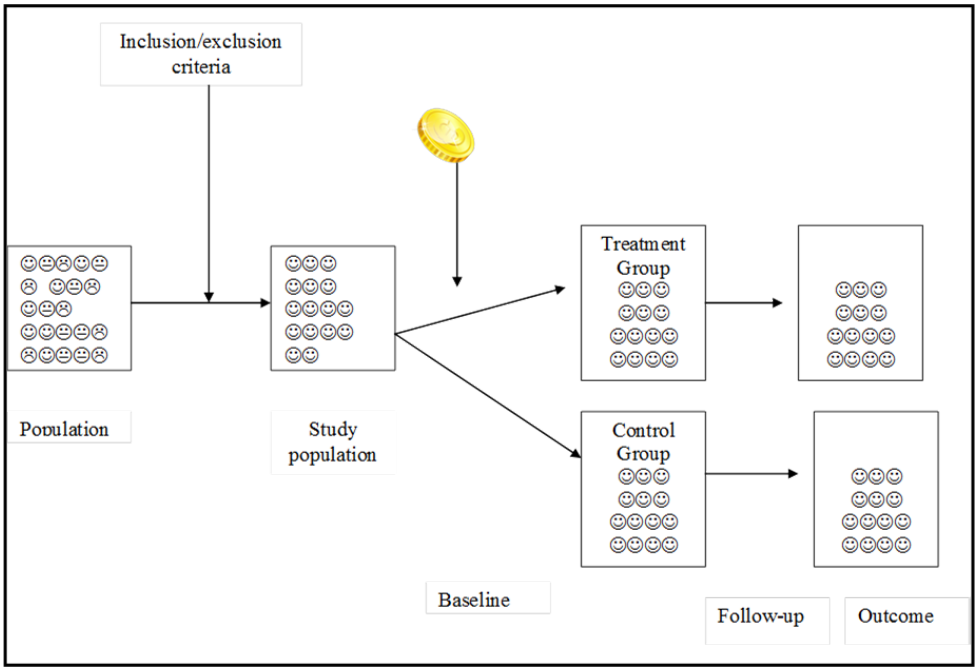

Figure 1. Diagram of the Basic Elements of a Randomized Controlled Trial

However, RCTs have several drawbacks. Their high cost means that they are often of short duration and underpowered to detect rare but serious side-effects or problems that arise only after prolonged use. The exclusion criteria typically result in concerns about generalizability. For example, RCTs involving medications for chronic obstructive pulmonary disease (COPD) often exclude people who also have asthma or other serious comorbidities, which excludes approximately 40% of people with COPD in Ontario.

Lastly, some studies simply cannot be conducted as an RCT. Once a treatment becomes the accepted standard of care, it can be difficult to ethically randomize people to receive no treatment; and conversely, it would not be ethical to randomize people to use alcohol or to smoke.

OBSERVATIONAL DATA

Observational data are data produced in circumstances where the choice of treatment is not under the control of the researcher. The researcher is relegated to the role of an observer, someone who can observe which treatment each person received, but cannot randomize subjects to one treatment or

another. In health services research there are two common sources of observational data: chart reviews (abstracting data from patient charts in a hospital, physician office, or other health care facility) and administrative data (data collected for other purposes, usually in order to make payments). Chart reviews can provide a rich set of clinical variables, and collecting observational data from charts can be relatively inexpensive if the study is restricted to a single facility or if the information can be obtained from electronic medical records. In the absence of electronic medical records, obtaining data from a large number of facilities using chart reviews becomes expensive and time-consuming.

Administrative data lacks the depth of clinical information found in charts. On the other hand, these data have already been collected (although data cleaning and linkage can still take a lot of work!), cover all of the care provided by the insurer, including both hospital and ambulatory care, and include information on all of the health care recipients and providers covered by the insurer.

Studies based on observational data are attractive whenever it is not feasible to use a controlled experiment, whether because it is considered to be unethical to withhold treatment, because the exposure to be studied is believed to be harmful, because patients will not agree to be randomized, or because an RCT would be too expensive. Observational studies are also extremely useful tools for hypothesis generating. Because the data cover a wide variety of subjects and providers, study results are likely to be generalizable. Furthermore, if the databases include a large number of individuals, they allow the study of rare outcomes.

The propensity score methods will be illustrated using an example of a study comparing two medications used to treat individuals with chronic obstructive pulmonary disease (COPD) (Gershon et al., 2011). The two medications are long-acting anticholinergics (LAACs) and long-acting beta agonists (LABAs). Both medications expand the airways in the lungs, making it easier for people with COPD to breathe. The short-acting forms of these medications have been associated with an increased risk of cardiovascular disease, but at the time of the study there was little information about the risks associated with the long-acting forms of the medications. The study was undertaken to compare the risk of cardiovascular events (e.g., heart attack, stroke) for the two medications.

While this paper uses a study comparing two medications to illustrate propensity score methods, the methods can be applied to the comparison of any two treatments, including medical vs. surgical treatment, different types of surgery, and treatment vs. no treatment or to compare the results of two non-medical programs.

Ontario is the largest province by population in Canada, with a population of 13 million people. A single payer, the Ontario Health Insurance Plan, covers the costs of all medically necessary services, and, for individuals aged 65 years and older, the cost of prescription drugs. The administrative databases collect information associated with health care services in order to make payments to pharmacists and physicians and to set hospital budgets. These datasets are linkable using unique, encoded identifiers and were analyzed at the Institute for Clinical Evaluative Sciences.

Figure 2 summarizes the main databases used for this study. Starting in the middle of the diagram, the Ontario Drug Benefit database contains one record for each prescription paid for by the plan, and includes the patient identifier, the type of medication, and the date the prescription was filled. The database was used to identify a cohort consisting of everyone who filled a new prescription for either a LAAC or a LABA during the study period. The baseline date was the date the prescription was filled. On the left side of the diagram are three databases which were searched in order to obtain the baseline characteristics of each person in the study. They provide information on characteristics such as age, sex, neighborhood socioeconomic status, and comorbidities reported prior to the baseline date. The Hospital Discharge database, shown on the right, was used to identify outcomes. This database contains a record of each admission to an Ontario hospital, including the patient identifier, and date and reason for admission. This allowed us to determine whether any of the individuals in the study cohort were hospitalized for a cardiovascular event, and if so, the time from baseline until hospital admission or death. Observations were censored if no hospitalization/death had occurred in the period covered by the data or if the individual emigrated and was no longer covered by the Ontario insurance plan.

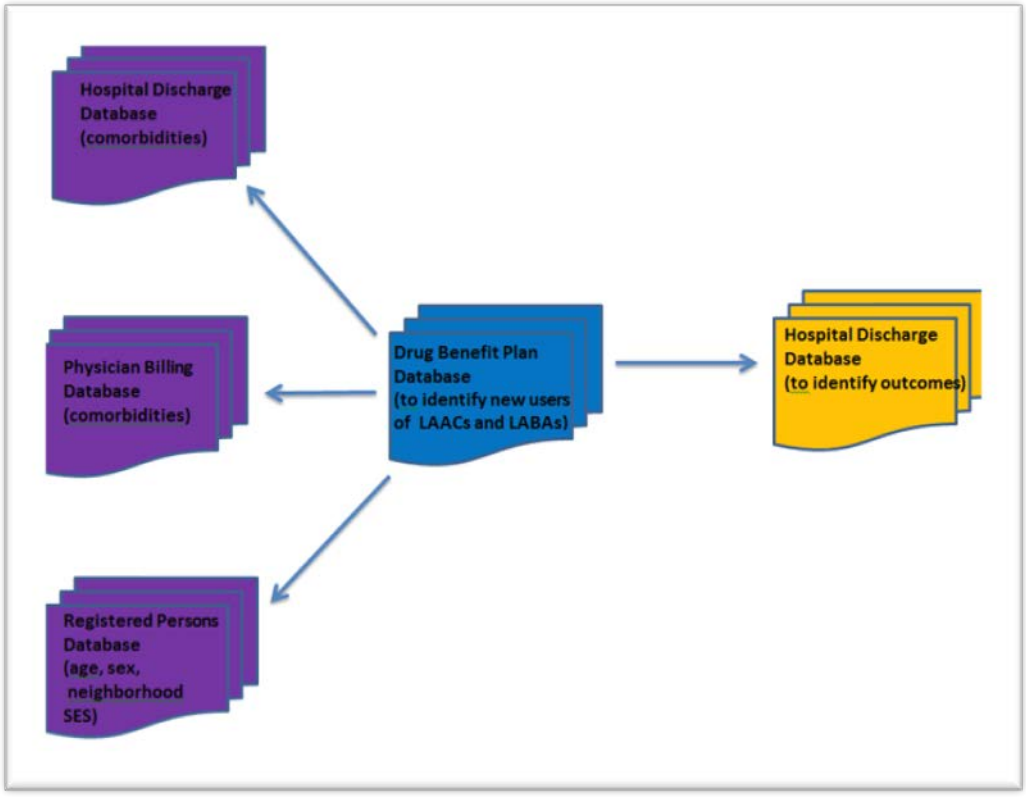

Figure 2. Administrative databases

STATISTICAL ADJUSTMENT FOR OBSERVATIONAL DATA

In the analysis of observational data, the concern is that patients who received one treatment differ from those who received the other treatment. They may be sicker, less compliant with their treatment, cared for by less skilled physicians, or have less access to care. The goal of the analysis is to use the available data in order to adjust for any differences that exist, so that the two treatments can be compared fairly.

Four statistical methods are available to adjust for differences between the two treatment groups:

1. Regression adjustment
2. Stratification
3. Propensity score methods
4. Instrumental variable analysis (not discussed in this paper).

Regression analysis asks if there is evidence of a difference in survival between the two drug groups, after adjusting for other risk factors present at baseline. It suffers from several drawbacks: it is difficult to assess whether the model has sufficiently adjusted for differences between the two treatment groups; when the outcome is rare there are limitations on the number of covariates that can be adjusted for; unlike an RCT, the results of the analysis estimate the conditional effect rather than the population-average (or marginal) effect; and the analysis is not separated from the study design, meaning that it is possible to keep tinkering with the analysis until you obtain the “right” answer.

In a stratified analysis, subjects are grouped into strata according to their baseline covariates. The two treatment groups are then compared within each stratum. Cochran (1968) demonstrated that five strata are often sufficient to remove 90% of the bias due to the variable used to form the strata. However, as the number of covariates increases, the number of strata increases, quickly resulting in strata that are too small to contain a useful number of subjects from both treatment groups.

Propensity score methods build on the concept of stratification. Rosenbaum and Rubin (1984) provided the insight that the bias from a set of baseline covariates can be eliminated by creating a single continuous variable which is a function of those covariates – the propensity score – and forming strata using this single variable.

THE PROPENSITY SCORE

The propensity score is the probability of receiving one of the two treatments, given the observed baseline covariates. Thus, it is easily calculated using the LOGISTIC procedure:

```
proc logistic data = laac_laba descending;
  model treatment_laac = age sex diabetes asthma ...;
  output out = propensity_score (keep = subject_id PS) p = PS;
run;
```

The predicted values obtained from the logistic regression represent the predicted probability that an individual was prescribed a LAAC rather than a LABA. It can be thought of as measuring the propensity of physicians to prescribe one drug rather than the other depending on the patient characteristics. Patients who fit the description of those likely to be prescribed a LAAC will have a high propensity score (even if they actually received a LABA). Those who have characteristics which, according to the logistic regression, make them likely to be prescribed a LABA will have a low propensity score (even if they actually received a LAAC).

Which covariates should be included in the logistic regression? The databases provide both variables believed to be associated with the choice of treatment and variables believed to be associated with the outcome. Variables affecting the outcome should always be included in the propensity score model, and there is evidence that including variables that affect only the choice of treatment will increase the variance of the final estimate of the treatment effect. (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006). However, in practice, it can be hard to distinguish between the two types of variables.

Datasets obtained using observational data can contain thousands of observations, allowing a non-parsimonious model to be fit (that is, all of the covariates are left in the model, regardless of their p-value). If overfitting is a concern, non-significant covariates can be removed.

Once the propensity score has been calculated, there are four ways to use it:

1. Covariate adjustment
2. Stratification
3. Matching on the propensity score
4. Inverse probability of treatment weighting using the propensity score

Covariate adjustment uses the propensity score itself as a covariate in the final analysis. For example, in our drug study, the survival analysis of time to hospitalization/death would include two predictors: the choice of medication and the propensity score. This method is not generally used because it relies on the values of the propensity score being correct and having a linear relationship with the outcome, and on the absence of an interaction between the propensity score and the treatment.

Early studies used the propensity score as stratification variable. The observations were divided into five strata, based on the quintiles of the propensity score. Most of the people in the quintile with the lowest scores will have been prescribed a LABA, but even in this quintile, there are likely to still be some people who were prescribed a LAAC. Similarly, most of the people in the quintile with the highest scores will have been prescribed a LAAC, but the quintile will still contain some who received a LABA. Outcomes for the two treatments are compared within each quintile, and then the 5 estimates are combined to get a single overall estimate of the difference between the two treatments.

MATCHING ON THE PROPENSITY SCORE

Empirical and theoretical research has shown that matching on the propensity score eliminates a greater

amount of treatment selection bias than does stratification.[4] A matched analysis tries to match each person in one treatment group with someone from the other group who has the same or similar propensity score. For matching, the logit of the propensity score is used, and two individuals are declared to be a match as long as their logit-transformed scores are within a "caliper" of one another. The usual value for the caliper is plus or minus 0.2 standard deviations of the transformed score.

A strength of a matched analysis is that it is possible to assess how well the matching achieved balance between the two matched treatment groups with respect to all of the measured covariates. If the two groups are not adequately balanced, additional terms, such as interactions and higher order polynomials for the covariates that remain unbalanced are added to the logistic regression equation, and the propensity score is recalculated. The matching process is then repeated, and balance is reassessed. This process is repeated until adequate balance is achieved.

Balance is assessed using standardized differences. Significance testing is not used. On one hand, significance testing is confounded with sample size. The matched groups, being smaller than the original sample, may not show "significant" differences due solely to the smaller sample size. On the other hand, samples based on administrative data are likely to be large enough so that even clinically meaningless differences are still statistically significant. The standardized difference is calculated as the difference between the two treatment groups divided by the pooled standard deviation of the difference.

For a continuous baseline variable, the standardized difference is:

$$d = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2 + s_1^2}{2}}}$$

For a binary categorical baseline variable, the standardized difference is:

$$d = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)}{2}}}$$

Often, the standardized difference is multiplied by 100 and expressed as a percentage. While there is no uniformly agreed on criterion for assessing standardized differences, many authors use a threshold of 10%.[4]

Table 1 illustrates the effectiveness of matching by presenting the results for a small subset of the covariates used to calculate the propensity score for our drug study. Before matching, there were large differences between the groups related to their receipt of specialist care, and use of other COPD medication. Matching reduced the total number of observations – there were observations in both treatment groups who could not be matched. However, matching produced two well-balanced treatment groups, even for those covariates which started out badly balanced.

	Before matching			After matching		
Covariate	LAAC N = 28,563	LABA N = 17,840	Stdized diff (%)	LAAC N = 15,532	LABA N = 15,532	Stdized diff (%)
Comorbidities Congestive heart failure	40.2	38.2	4.1	39.0	39.2	0.4
COPD care Previous spirometry	69.7	74.3	10.2	72.4	73.0	1.3
COPD care One or more visits to a specialist in the previous year (%)	44.7	50.5	11.6	49.0	49.1	0.2
Medication Short-acting cholinergic (%)	54.3	45.7	17.4	49.6	49.3	0.6

Table 1. Comparison of treatment differences before and after matching for a subset of covariates

The final analysis of the matched treatment groups must take the matching into account. For example, if the outcome variable is a continuous variable, a matched t-test should be used; for proportions, McNemar's test will take the matching into account. Our drug study involved time-to-event with censored data, and we used survival analysis, stratifying on the matched pairs.

By analyzing only the matched pairs, the analysis ensures that, like an RCT, only patients who are potential candidates for both treatments are included in the final comparison. Matching results in pairs of patients who are similar with respect to the available covariates; and by using the propensity score, we were able to match simultaneously on a large number of potential confounders. Individuals who were not matched are excluded from the final analysis, and therefore the resulting matched sample may not be representative of the entire patient population.

INVERSE PROBABILITY OF TREATMENT WEIGHTING USING THE PROPENSITY SCORE

It may be helpful to think of inverse probability of treatment weighting (IPTW) as a missing data problem. We'd like to compare each person's outcome while taking a LAAC with their outcome while taking a LABA. Unfortunately, for those who were prescribed a LAAC, information on how they'd have done on a LABA is missing. And for those who were prescribed a LABA, information on how they'd have done on a LAAC is missing. The missing information is provided by finding a similar person in the opposite treatment group, and using their outcome to impute the missing information.

Alternatively, inverse probability of treatment weighting can be thought of as being similar to a survey sampling problem, in which some groups of people are oversampled, while others are undersampled, and then weights are used to make inferences about the population.

The weights for each observation are $w_i = \frac{Z_i}{PS} + \frac{1-Z_i}{1-PS}$ where Z is an indicator variable, equal to 1 if the subject was treated and 0 otherwise (or equal to 1 for the treatment of interest and 0 for the reference treatment).

IPTW analysis creates two datasets, equal in size to the original dataset. One will be used to estimate the outcome for the treated group and the other will be used to estimate the outcome for the untreated group. All of the observations are included in both datasets. Thus, for those people who were treated, their outcome in the untreated dataset is missing, while for those people who were untreated, their outcome in the treated dataset is missing.

Table 2 shows the contribution of five imaginary subjects to the treatment dataset. Subjects 1, 2, and 5 received the treatment, and therefore we know what the effect of treatment was for them. Subjects 3 and 4 were not treated. They do have outcomes (Y_3 and Y_4), but those outcomes were not obtained during treatment, and we do not know what the effect of treatment was for them.

Subject	Z	Propensity Score	Weight (w)	Outcome	Effect of treatment
1	1	0.2	5	Y_1	Y_1
2	1	0.5	2	Y_2	Y_2
3	0	0.3	1.4	Y_3	missing
4	0	0.4	1.7	Y_4	missing
5	1	0.8	1.3	Y_5	Y_5
...					
N					

Table 2. Illustration of the creation of a dataset to calculate the average outcome if all subjects were treated

To use this dataset to estimate the outcome with treatment, we calculate the average outcome with treatment as $\frac{1}{N}\sum_{i=1}^{N} w_i Z_i Y_i$. Since Z_i is equal to 0 for those who were not treated, they do not contribute to the estimate of the average outcome of the treatment.

Table 3 shows the same five observations, but this time, it shows their contribution to the untreated dataset.

Subject	Z	Propensity Score	Weight (w)	Outcome	Effect of no treatment
1	1	0.2	5	Y_1	missing
2	1	0.5	2	Y_2	missing
3	0	0.3	1.4	Y_3	Y_3
4	0	0.4	1.7	Y_4	Y_4
5	1	0.8	1.3	Y_5	missing
...					
N					

Table 3. Illustration of the creation of a dataset to calculate the average outcome if all subjects were untreated

Using this dataset, the estimated average outcome without treatment is $\sum_{i=1}^{N} w_i \ (1 - \ Z_i) \ Y_i$.

The treatment effect is obtained by taking the difference between the estimated outcome when treated and the estimated outcome when untreated.

Alternatively, an inversely weighted analysis can be viewed as being similar to the analysis of a survey such as the National Health Interview Survey in the U.S. or the Canadian Community Health Survey in Canada. Both of these surveys use a complex sampling strategy involving stratification, clustering, and oversampling of smaller subpopulations. Sampling weights are assigned to each respondent, reflecting the probability that this respondent was selected for inclusion in the study. By applying the weights to each subject's responses, the responses can be combined to produce an estimate which reflects the results for the entire population.

With this in mind, we can view table 2 as a survey, with weights assigned to each Subject to reflect the probability of that person being in their treatment group. Subject 1 received the treatment. However, as indicated by the low propensity score, this person had a low (20%) chance of being prescribed the treatment. There must be 4 people who are similar to Subject 1 but who were untreated – thus, their outcome under treatment is missing. However, we expect that their outcome would have been similar to that of Subject 1, since their characteristics are similar. Therefore, when the treatment outcome is calculated, Subject 1 is assigned a weight of 5, indicating that this person represents herself plus 4 people from the untreated group.

Subject 2, who also received the treatment, had a higher (50%) probability of receiving the treatment. Therefore, there should be only 1 person in the untreated group who looks like Subject 2. Subject 2's weight is 2, indicating that in the calculation of the treatment outcome, she represents herself plus one additional person.

IT'S NOT QUITE MAGIC

Propensity score methods provide a useful way of reducing the dimensionality of an analysis, by using a single measure, the propensity score, to capture the information contained in a large number of covariates. This can be very useful, especially when studying a rare outcome. If the propensity score is used to match individuals from the two treatment groups, standardized differences can be used to demonstrate that the two matched groups are similar on all of the measured covariates. However, there are drawbacks. While propensity score methods can remove biases related to the measured covariates, they remove hidden biases only to the extent that the unmeasured variables are correlated with the variables available to compute the score. Administrative databases, in particular, are certain to be missing key covariates (e.g., smoking history, living arrangements).

Another drawback to observational data is that it may be difficult to define the baseline time point. Analogous to an RCT, this should be the time at which the treatment decision was made, but that time often cannot be determined from the data.

Observational data do not provide information about patient compliance or about the quality of follow-up care. Whereas RCTs typically measure the efficacy of a treatment, or its benefits under ideal conditions, observational studies examine the effectiveness of the treatment, or its benefits under routine "real world" conditions.

Like an RCT, a propensity score-based study does not predict the outcome for a person with a given set

of characteristics. The study does not look at the role of covariates other than treatment choice in predicting the outcome, nor does it examine who will benefit most from a given treatment. Both RCTs and propensity score-based analyses estimate the impact of moving a population from one treatment to another (or from no treatment to treatment).

And finally, propensity score methods do not eliminate the need to think carefully about cohort selection and outcome definition.

CONCLUSION

A wealth of information is contained in observational data, available to answer questions in health services research, psychology, the social sciences, and economics. Observational data allow us to study questions which cannot otherwise be addressed, whether for ethical or logistical reasons. They allow us to search for rare outcomes, and to follow subjects over long periods of time. However, there are valid concerns about treatment bias. Propensity score methods provide one set of tools to address these concerns, although the researcher must still think carefully about possible sources of unmeasured confounding and consider how these may affect the results.

REFERENCES

Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med. 2007; 26: 734-53.

Brookhart, M. A., Schneeweiss, S, Rothman, K. J, Glynn, R. J, Avorn, J, Stürmer, T. (2006) Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149-56.

Cochran, W. G (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295-313.

Gershon, A., Croxford, R., To, T., Stanbrook, M. B., Upshur, R., Sanchez-Romeu, P. & Stukel, T. (2011). Comparison of inhaled long-acting β-agonist and anticholinergic effectiveness in older patients with chronic obstructive pulmonary disease: a cohort study. *Annals of Internal Medicine, 154(9)*, 583-92.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association,79*, 516-525.

ACKNOWLEDGMENTS

The research used to illustrate propensity score methods was *supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred.* Parts of this material are based on data and information compiled and provided by the Canadian Institute for Health Information (CIHI). The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources and CIHI. We thank IMS Brogan Inc. for use of their Drug Information Database.

RECOMMENDED READING AND ADDITIONAL RESOURCES

- Austin PC. A tutorial and case study in propensity score analysis: An application to estimating the effectof in-hospital smoking cessation counseling on mortality. Multivar Behav Res. 2011; 46: 119-151.
- Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011; 46: 399-424.
- Matching can be performed using the GMATCH macro, available from http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-

biomedical-statistics-informatics/software/locally-written-sas-macros

- For a macro that calculates standardized differences see SAS Global Forum paper 335-2012: D Yang and JE Dalton. A unified approach to measuring the effect size between two groups using SAS®, available at http://support.sas.com/resources/papers/proceedings12/335-2012.pdf. A copy of the macro can be downloaded from http://www.lerner.ccf.org/qhs/software/lib/stddiff.sas

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ruth Croxford
Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada
ruth.croxford@ices.on.ca
www.ices.on.ca