ISSUE FOUR July 2014

Evidence Notes

Bridge Medical
Evidence that matters

Handling non-randomised data – Part 1: Propensity Scoring

Welcome to this new edition of Evidence Notes. Our aim with these newsletters is to write short, informative articles on a range of topics in the evidence space. With all the current discussion and debate around "real world evidence", we return to the age-old question on how to ameliorate the challenges of bias and confounding in non-randomised data sets. In this edition we describe approaches which address bias associated with *known* confounders i.e. multiple regression and propensity scoring, with a particular focus on propensity scoring. We describe these approaches *without* complex statistical terminology or equations – the aim of this piece is simply to give our readers some idea of when different techniques might be applicable in different circumstances, and some of the key drawbacks. The next edition of Evidence Notes will describe approaches that address bias associated with unknown or unmeasured confounders i.e. "instrumental variables".

Randomised controlled trials (RCT's) are the gold standard approach to study design. If any confounders are present that may influence outcome, whether known or unknown, these are likely to be balanced randomly between groups. However, randomisation is not always possible or desirable, especially when the goal of research may be to "observe" the "real world".

Whilst observational research offers many benefits over RCT designs they are more prone to bias and confounding. Investigators may influence treatment assignment and therefore direct comparisons of outcomes from the treatment groups may be misleading. For example, comparing the effect of different interventions on outcomes across subject groups that may have different baseline parameters (such as severity of illness or age or gender) is prone to significant confounding.

There are though statistical approaches that can be used in observational research to limit the potential impact of confounders on the outcome of interest. The best known approach is multiple regression analysis. This article will highlight some of the limitations of regression analysis, and highlight the potential role of an additional approach – propensity scoring (PS).

Figure 1 – Graphical Illustration of Statistical Approaches Used in Observational Research to Handle Known Confounding

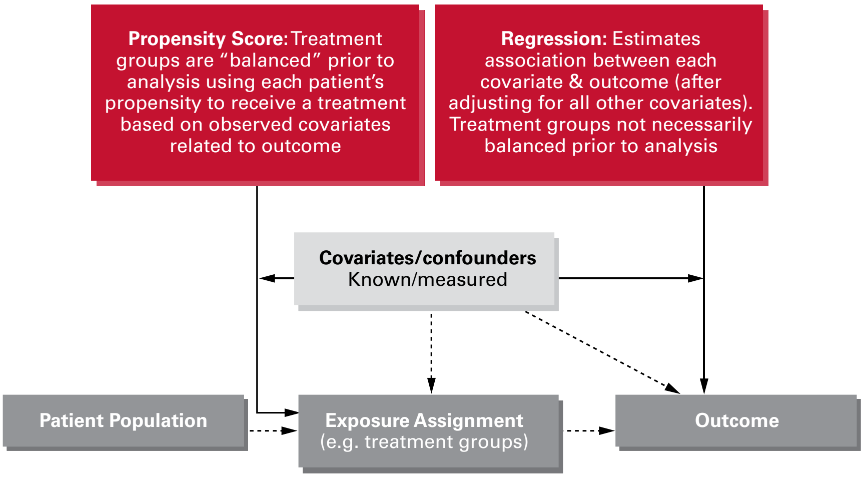

Propensity Score: Treatment groups are "balanced" prior to analysis using each patient's propensity to receive a treatment based on observed covariates related to outcome
Regression: Estimates association between each covariate & outcome (after adjusting for all other covariates). Treatment groups not necessarily balanced prior to analysis
Covariates/confounders Known/measured
Patient Population
Exposure Assignment (e.g. treatment groups)
Outcome

Figure 1 provides a simple illustration for the respective roles of regression and PS.

Multiple regression is the most commonly used statistical technique to overcome bias in observational studies. It is an approach which accepts that the groups are imbalanced and tries to minimise this by adjusting for each confounding factor, leaving only the variation linked to a single explanatory factor e.g. treatment. Limitations with this regression analysis include:

1. It should not be used where there are a large number of variables and rare outcomes (~8-10 outcome events per variable have been recommended for multivariable regression models)
2. Studies are typically designed (powered) to assess the effects of a single factor, rather than a single factor in the context of many other factors. Some of the assumptions in the design are therefore prone to error and, since in regression analyses these are not necessarily balanced across groups, they may impact the ability to statistically detect effects.
3. It does not take into account confounders which are either unknown/unobserved or unavailable (e.g. in administrative databases). This may lead to bias and error especially where the magnitude of effect on outcome is weak or modest.

In contrast, PS methods can address the first two issues, although they are also unable to deal with unobserved or "missing" confounders. This technique was first used by Rosenbaum et al. (1983) and has since been adopted in a variety of fields including epidemiology, health services research, economics and social sciences. Although PS methods can be used in a variety of ways, the focus of the current article is on its specific use in treatment intervention studies, including prospective observational studies or registries, or a retrospective investigation of existing databases.

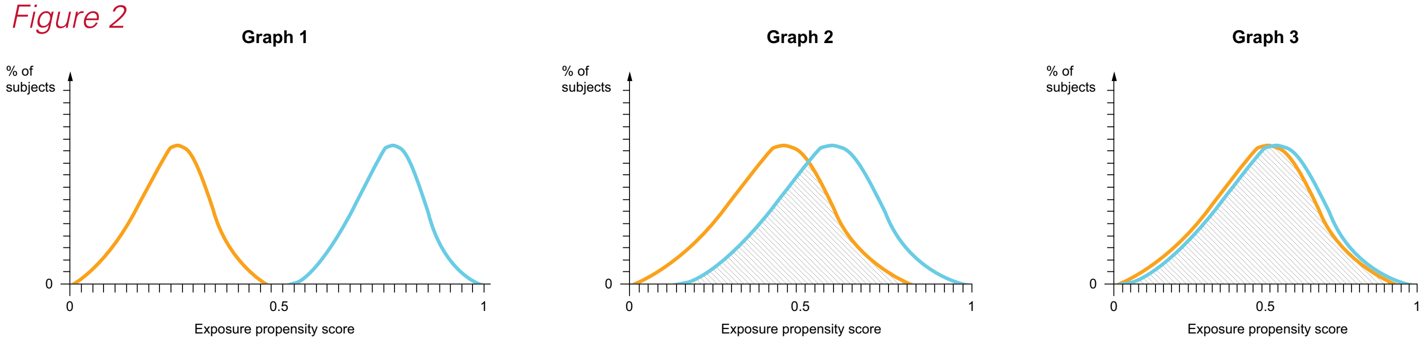
The key features of propensity scoring are as follows (also see *Figure 2* for a graphical illustration of unmatched data v matched data v randomised data)

- In PS approaches, subjects with the same *propensity* to receive treatment (based on the patients' baseline characteristics) are selected for comparison. This is how the balance between treatment groups is created and is the key difference from regression-based approaches. This is important because in a typical observational study treatments are not assigned randomly and groups (treated and untreated) may systematically differ at baseline.
- The propensity score is the probability (represented by a single score between 0 and 1) of receiving a treatment based on those known covariates believed to be related to outcome. In an RCT, where assignment to treatment is random, the probability of receiving one or other of the treatments would be 0.5.
- Estimating PS can be done in several ways but, most commonly, multivariable logistic regression models are used which include all baseline patient characteristics as well as any clinically relevant interactions.
- PS-methods are generally comparable with results from RCTs across a wide variety of indications and outcomes. It has been suggested by some experts that matching on PS may often result in a better balance of variables than can be obtained via randomisation.
- Matching subjects on the basis of PS can clearly identify subjects with little overlap on covariates, and these can be excluded from the analysis, whereas these differences might be obscured in regression analyses.

Table 1: Case Study using Propensity Scoring

Study Type	German Stroke Registry; retrospective
Population	Population Patients with ischaemic stroke in centres performing tissue plasminogen activator (t-PA) therapy; N = 6,269
Background	Observational studies have shown increased risk of death associated with t-PA treatment in these patients; RCTs have shown no causal association between t-PA treatment and death
Aim	Compare different analyses to adjust for confounding on the effect of t-PA on deaths following ischaemic stroke
Key Findings	• Unadjusted odds ratio (OR) for t-PA treatment & death after ischemic stroke was 3.35 (95% confidence interval (CI): 2.28, 4.91) vs 1.17 for propensity-matched subjects (95% CI: 0.68, 2.00) and vs 1.93 (95% CI: 1.22, 3.06) for logistic regression without PS [NB pooled relative risk in meta-analysis of several RCTs was 1.16 (95% CI: 0.95, 1.43)*] • For treated patients with a low propensity score, risk of dying was high. In patients with PS ≥ 0.05 (i.e. those perhaps less likely for treatment to be contraindicated) the estimated OR for all methods did not significantly differ from 1 or from the results of RCTs.
Key Findings	In contrast to findings from unadjusted observational studies, the propensity matched estimate showed no statistically significant association between t-PA treatment and death and was very similar to risk estimates obtained from RCTs. The propensity method was also able to identify a population of treated patients with a low propensity for t-PA (i.e. potentially contraindicated for use) in whom death rates were high.

Figure 2



The orange and blue lines represent the proportion of patients treated with interventions A & B, respectively, as a function of their propensity for treatment assignment based on observed covariates related to outcome. Using the propensity score, subjects from each treatment group can then be matched (the hatched areas represent the most simple matching using a 1:1 ratio of case:control) and their outcomes compared between treatments.

Graph 1 – shows two populations with very different propensities for treatment assignment (non-matched)
Graph 2 – shows populations with 1:1 matched (hatched) propensities for treatment assignment
Graph 3 – shows a typical RCT population with overlapping distributions for both treatments and a propensity score of 0.5

A case study where PS matching was employed is provided in *Table 1*. In this case, the application of PS led to a more accurate clinical interpretation of the available data. For further examples please refer to Borah (2014) and Heinze (2011).

There are, however, issues related to the us of PS matching and some of the pros and cons are summarised in *Table 2*.

In summary, since the likelihood of receiving an intervention is based on observed covariates, PS methods are useful for adjusting for these known confounders and in balancing the population prior to analysis may offer some advantages to multiple regression approaches.

It is not always possible to know or measure all potential confounders, and in these circumstances other approaches – such as Instrumental Variables – may be required. This topic will be the subject of the next article.

Further information on Propensity Scoring can be found in the references provided.

Table 2: Summary of pros & cons

Pros	Cons
Provides balance in the comparison groups akin to that in RCTs	PS only accounts for measured covariates; does not account for unmeasured (or hidden) covariates
Can compare outcomes (& allows causal inference) in those with similar PS in different treatment groups	Some datasets (especially database studies) may not record all the variables of interest
PS better than regression when there are few outcomes and a large number of variables	Additional sensitivity analyses are recommended
PS matching identifies subjects with little overlap on covariates & can be excluded from the analysis; these differences might be obscured in regression analyses	PS in treated and untreated groups may not overlap - estimation of treatment effect resides only in those whose PS overlap
Easier to assess the degree of overlap of baseline covariates in PS vs regression approaches	May not be including all potentially useful data that are representative of "treatment" in the real world.

Martin Jones, Aiden Flynn and Paul Gandhi

Corresponding author:
martinjones@bridgemedical.org

References

1. Armstrong K. Methods in comparative effectiveness research. J Clin Oncol. 2012 Dec 1;30(34):4208-14. http://www.ncbi.nlm.nih.gov/pubmed/23071240
2. Johnson ML, Crown W, Martin BC, et al. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report--Part III. Value Health. 2009 Nov-Dec;12(8):1062-73. http://www.ncbi.nlm.nih.gov/pubmed/19793071
3. Normand SL, Sykora K, Li P, et al. Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. BMJ. 2005 Apr 30;330(7498):1021-3. http://www.ncbi.nlm.nih.gov/pubmed/15860831
4. Normand SL. Some old and some new statistical tools for outcomes research. Circulation. 2008 Aug 19;118(8):872-84. http://www.ncbi.nlm.nih.gov/pubmed/18711024
5. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996 Dec;49(12):1373-9. http://www.ncbi.nlm.nih.gov/pubmed/8970487
6. Cepeda MS, Boston R, Farrar JT et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol. 2003 Aug 1;158(3):280-7. http://www.ncbi.nlm.nih.gov/pubmed/12882951
7. Rosenbaum, PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41-55 http://biomet.oxfordjournals.org/content/70/1/41.abstract
8. 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 Oct 15;17(19):2265-81. http://www.ncbi.nlm.nih.gov/pubmed/9802183
9. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med. 1997 Oct 15;127(8 Pt 2):757-63. http://www.ncbi.nlm.nih.gov/pubmed/9382394
10. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011a May;46(3):399-424. http://www.ncbi.nlm.nih.gov/pubmed/21818162
11. Austin PC. A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. Multivariate Behav Res. 2011b;46(1):119-151. http://www.ncbi.nlm.nih.gov/pubmed/22287812
12. Alemayehu D, Alvir JM, Jones B, et al. Statistical issues with the analysis of nonrandomized studies in comparative effectiveness research. J Manag Care Pharm. 2011 Nov-Dec;17(9 Suppl A):S22-6. http://www.ncbi.nlm.nih.gov/pubmed/22074671
13. Borah BJ, Moriarty JP, Crown WH, et al. Applications of propensity score methods in observational comparative effectiveness and safety research: where have we come and where should we go? J Comp Eff Res. 2014 Jan;3(1):63-78. http://www.ncbi.nlm.nih.gov/pubmed/24266593
14. Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. Eur Heart J. 2011 Jul;32(14):1704-8. http://www.ncbi.nlm.nih.gov/pubmed/21362706
15. Ahmed A, Husain A, Love TE, et al. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. Eur Heart J. 2006 Jun;27(12):1431-9. http://www.ncbi.nlm.nih.gov/pubmed/16709595
16. Schneeweiss S, Gagne JJ, Glynn RJ, et al. Assessing the comparative effectiveness of newly marketed medications: methodological challenges and implications for drug development. Clin Pharmacol Ther. 2011 Dec;90(6):777-90. http://www.ncbi.nlm.nih.gov/pubmed/22048230
17. Kuss O, Legler T, Börgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. J Clin Epidemiol. 2011 Oct;64(10):1076-84. http://www.ncbi.nlm.nih.gov/pubmed/21482068
18. Lonjon G, Boutron I, Trinquart L, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. Ann Surg. 2014 Jan;259(1):18-25. http://www.ncbi.nlm.nih.gov/pubmed/24096758
19. Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. Eur Heart J. 2012 Aug;33(15):1893-901. http://www.ncbi.nlm.nih.gov/pubmed/22711507
20. Collins GS, Le Manach Y. Comparing treatment effects between propensity scores and randomized controlled trials: improving conduct and reporting. Eur Heart J. 2012 Aug;33(15):1867-9. http://www.ncbi.nlm.nih.gov/pubmed/22745354
21. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol. 2006 Mar;98(3):253-9. http://www.ncbi.nlm.nih.gov/pubmed/16611199
22. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol. 2006 Feb 1;163(3):262-70. http://www.ncbi.nlm.nih.gov/pubmed/16371515
23. Wardlaw JM, Sandercock PA, Berge E. Thrombolytic therapy with recombinant tissue plasminogen activator for acute ischemic stroke: where do we go from here? A cumulative meta-analysis. Stroke. 2003 Jun;34(6):1437-42. http://www.ncbi.nlm.nih.gov/pubmed/12730560