BMJ

BMJ 2013;347:f6409 doi: 10.1136/bmj.f6409 (Published 11 November 2013)

RESEARCH METHODS & REPORTING

Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research

Propensity score based methods are used increasingly to evaluate the effectiveness of treatments when evidence from randomised trials is not available. However, users need to be aware of their strengths and limitations

Nick Freemantle *professor of clinical epidemiology and biostatistics*[1 2], Louise Marston *senior research statistician*[1 2], Kate Walters *senior clinical lecturer in primary care and epidemiology*[1], John Wood *principal research statistician*[1 2], Matthew R Reynolds *director, economics and quality of life research*[3 4], Irene Petersen *reader in epidemiology and statistics*[1 2]

[1]Department of Primary Care and Population Health, UCL Medical School (Royal Free Campus), London NW3 2PF, UK; [2]PRIMENT Clinical Trials Unit, UCL Medical School ; [3]Lahey Clinic Medical Center, Burlington, MA, USA ; [4]Harvard Clinical Research Institute, Boston, MA, USA

For well rehearsed reasons, randomised trials are established as the mainstay of the evaluation of healthcare interventions. Indeed, as far back as 1935 Ronald Fisher commented that, "the simple act of randomisation assures the internal validity of the test for significance,"[1] before going on to lambast Charles Darwin for making strong conclusions from observational data. An irascible eugenicist and misogynist, Fisher was a brilliant but flawed genius; but we ignore his guidance on avoiding bias at our peril.

Though Fisher's aphorism remains true today, it addresses only part of the challenge. Decision makers could be equally interested in the external validity of a research finding, often asking for information about the effectiveness of treatments in the real world. Randomised trials, particularly those undertaken to support an application for marketing authorisation of a new medical product, may include by design only a stylised subset of patients with the particular condition: patients who are adherent and somewhat positively disposed to at least one of the treatment options (as identified by their agreement to be randomised); patients who are relatively lacking in comorbidities (to reduce the risk of serious adverse events that might confound the assessment of safety); and patients who are unrepresentatively young and often predominantly male (as investigators tend to be clinical specialists who draw from their local patient population, rather than generalists specialising in the complex problems of older people). Furthermore, participants are usually low risk because treatment is compared with a placebo (since the regulatory bodies require only proof that a product works not information on how it compares with existing treatments). Randomised trials are also expensive, with little change from $50m (£31m; €35m) for a landmark regulatory study, and they have a long lead time from inception to completion. This means that current patients with serious conditions may not benefit from the results even if the trials are conducted.

In attempts to address some of these limitations (trials conducted in the wrong or unrepresentative populations, or not done at all) researchers may turn to observational methods and the rich array of observational data to fill in some of the gaps.[2] In this paper we describe, through a series of examples, some of the potential advantages and perils of observational studies and suggest some strategies to negotiate the challenges safely.

Propensity score analyses

Propensity scores were described by Rosenbaum and Rubin in 1983[3] as a deft means of accounting for known confounders or biases in estimation. They have developed a central place in observational research, being used in many settings, including those that do not lend themselves to randomised trials.[4] Smeeth et al used propensity score based analyses to address confounding in the comparison of people treated and not treated with statins.[5] Taking data from the Health Improvement Network (THIN), a data set based on general practice records, they fitted a statistical model to estimate the individual likelihood that patients would be prescribed a statin, using a list of 39 potential explanatory variables unaffected by exposure to statins,

Correspondence to: nicholas.freemantle@ucl.ac.uk

including demographic and medical history, prescribed drugs, social deprivation, and consultation behaviour. This provided a propensity score for each patient (that is, the model's prediction of the likelihood of receiving a statin). They used these scores to adjust for confounding when estimating the differences in outcome between patients receiving and not receiving statins but with a similar propensity for receiving statins. We can have some confidence that the adjustment using propensity scores was successful because their results were closely in line with those for vascular outcomes in randomised trials, lending support to the notion that other non-vascular outcomes (the focus of their research) may also reasonably be compared.

Propensity scores can be used for adjustment in statistical models or to create matched groups by selecting treatment and control patients with similar propensity scores. In both applications, propensity scores are used to account for known confounders and their use may lead to quite different results from those gleaned from unadjusted comparisons. Indeed, propensity score models can be considered a special case of multivariable adjustment. Propensity matched analyses are particularly attractive as they include in the analyses only participants who have a similar propensity score and thus baseline characteristics. Matched propensity score evaluations also make it straightforward to compare the characteristics of treated and untreated groups and promote analysis strategies analogous to those used for randomised trials, although difficulties in achieving adequate matches may lead to small sample sizes and reduce external validity.

When things go awry

Although potentially helpful, the use of propensity scores does not assure the internal validity of the significance test, and decision makers need to be wary of making inferences from their results. This can be illustrated by the case of spironolactone, an aldosterone inhibitor that in the Randomized Aldactone Evaluation Study (RALES) reduced mortality in patients with severe heart failure (hazard ratio 0.70, 95% confidence interval 0.60 to 0.82, P<0.001).[6] The result was independently confirmed in two other trials.[7 8] Using a propensity score matching approach, we attempted to replicate the RALES trial[6] using data from the Health Improvement Network, with the ultimate objective of bridging from the trial population to a real world population of people with heart failure.[9] We included only patients who had recently been treated with high dose loop diuretics (≥80 mg furosemide a day or equivalent), which indicates congestion, and excluded patients on the palliative care register, those with renal dysfunction, and those with recent cancer or unstable angina, liver failure, or a heart transplant. We used a large number of indicators of patient demography, comorbidities, and drug treatments to develop a propensity score. This was used to make two tightly matched groups of patients (n=4412) treated and not treated with spironolactone. We also did many supportive analyses, essentially taking a series of different defensible approaches to creating the propensity scores in order to explore consistency, adding further potential risk factors such as recent acute medical hospital admission and increasing the required precision by which matches were acceptable.

Survival in the spironolactone treated groups in RALES[6] and in our propensity matched study was remarkably similar, with just over 80% survival in both cohorts at one year. However, when we compared the tightly matched propensity score groups, rather than reducing mortality, spironolactone seemed to be associated with a substantial increase in the risk of death (figure⇓). So must we conclude that spironolactone is dangerous in heart failure and should not be used, favouring the findings of the propensity matched analyses over those of the randomised clinical trials? Such strong conclusions have been drawn for other drugs on the basis of evidence of similar quality.[10 11] But we contend that such a conclusion would be quite unsafe. Below we explore some of the reasons why propensity score analyses may give incorrect answers.

Unknown bias

Randomisation, when properly conducted, avoids bias by distributing both known and unknown patient characteristics between the experimental conditions on the basis of the play of chance. This provides a good basis for comparison between the groups. It also underpins our statistical analyses because there are just two potential (orthogonal) explanations for any difference observed between the experimental groups: that the differences are due either to the randomised treatment or to the play of chance. If it is implausible that chance is responsible for the observed difference because the P value is very small or the confidence intervals are a long way from the point of no difference, it must be due to the effects of treatment. It is, as Fisher recognised, a neat trick.

Propensity score based analyses, by contrast, account only for known and observed patient characteristics. We hope that by balancing these known confounders we may derive an unbiased estimate of the effects of treatment. As Rosenbaum and Rubin pointed out in 1983,[3] this notion requires the assumption that treatment assignment (in our case spironolactone or no spironolactone) is "otherwise ignorable"—that is, that no additional unknown processes related to patient severity are associated with determining who will or will not receive treatment. Of course biases can be found in randomised trials and propensity score analyses, and both have the potential to be conducted poorly. For example, attribution bias in non-blinded trials, loss to follow-up, or failure to follow the intention to treat principle will all lead to biased results from randomised trials. In randomised trials and observational studies clinicians may introduce treatments considered in the best interest of the patient but which could undermine the intended comparison. For example, in the spironolactone study investigators may have introduced alternative treatments for heart failure that were not adequately recorded (although we found no evidence of this). In randomised trials the investigators may (correctly) introduce treatments that undermine the validity of the trial comparison, while acting in the best interests of the patient. Both trials and propensity score analyses must be conducted to high methodological standards, although ensuring this for propensity score analyses, which are intrinsically more complex, can be harder.

The design of a study can sometimes make the decision to treat a patient "otherwise ignorable"—for example, in the use of the propensity score to identify appropriate subjects to compare rhythm and rate control in an observational study based on the AFFIRM trial.[12 13] In the AFFIRM trial participants were randomised to rhythm control drugs or rate control drugs, but the actual drug prescribed in that class was determined by the investigator physician; the AFFIRM investigators were comparing treatment strategies (rate versus rhythm control) rather than individual drugs. However, Saksena and colleagues sought to use these data to make valid comparisons between particular antiarrhythmic drugs and the rate control strategy.[13] Although the doctor chose the antiarrhythmic—and so the choice

may carry information about severity—the decision not to use an antiarrhythmic drug is ignorable by design (because it was allocated on the play of chance). So similar comparator subjects should exist in the rate control group for each antiarrhythmic drug chosen, and propensity score matching should provide an excellent basis for identifying control subjects. In a truly observational setting there is the potential risk that the choice of treatment is driven by patient characteristics, resulting in few if any control subjects being available, which reduces estimation precision and external validity.

Confounding by indication

Confounding by indication is the situation where allocation to treatment is not otherwise ignorable but instead subject to some latent (unrecognised or unmeasured) process associated with those who are treated—for example, when skilled clinicians use their expert judgment to decide whether to treat a patient and this judgment includes criteria describing the severity of the condition or the frailty of the patient not included in the propensity score or, more likely, not even formally measured. The challenge of a latent function such as confounding by indication is that it (by definition) cannot be measured directly but only tangentially through its effects, if it is recognised at all.

One obvious way to assess the performance of a propensity score is to examine its performance for homogeneity at different points on the propensity score scale.[14] In the figure⇓ we describe the effect of randomisation to spironolactone rather than placebo in the RALES trial,[6] then provide estimates derived from our propensity score matched model and for the four quartiles of the propensity scores used to generate the matched comparison. The hazard ratio for the overall propensity score analysis differs from that in the RALES trial by 6.4 standard errors (P<0.001), but there is also substantial variation between different values on the propensity score scale (test for interaction between the propensity score and spironolactone P=0.003).

Although the matched comparisons performed poorly for all values of the propensity scores, they were particularly misleading for participants who scored below the median likelihood of receiving spironolactone. One explanation for the spironolactone effects across the range of the scale, and in particular in the apparently low propensity subjects, is that the prescriber making the clinical decision to treat used additional important information on severity of heart failure that the propensity score did not capture, and so the match was made with inappropriately low risk individuals. In other words, the decision to prescribe was not otherwise ignorable.

When is it helpful to use a propensity score analysis?

Had the evidence for spironolactone from the RALES trial been less convincing we may have been tempted to conclude that spironolactone was indeed associated with substantial harm, and if that finding had been listened to, patients may have died who otherwise would have lived longer. There are certainly cases where patient harm seems likely because of the interpretation of propensity score analyses that are open to the risk of confounding by indication.[10 11 15]

The salutary example of our propensity matched analysis comparing patients who were treated with spironolactone and those who were not in patients with moderate to severe heart failure illustrates the perils for the unwary but also helps us to consider some rather limited strategies to identify or circumnavigate the problems.

Firstly, as confounding by indication is not directly measurable, this offers challenges to the analyst. Propensity score analysis will not lead to biased estimates of treatment effect if it is used in situations where the treatment decision is otherwise ignorable. For example, the analysis of the effects of statins by Smeeth et al[5] examined a treatment that tends to be used algorithmically based on estimated long term risk of major morbidity or mortality, rather than in a response to observed patient morbidity, and so the treatment decision may be considered more likely to be otherwise ignorable.

Secondly, a useful precaution against unsafe inference from an observational study is to start with the replication of a known treatment effect and bridge from there to consider further, hitherto unanswered, questions. This was the approach taken in the analysis by Smeeth et al, who first replicated the vascular findings of the randomised trials before bridging to examine other outcomes of importance.[5] It was also our aim in the spironolactone analyses, where our intention was to bridge from a replication of the RALES trial to examine treatment effects in a broader group of patients, (women, those who are older, those with comorbidities and other different characteristics).

Thirdly, Rosenbaum and Rubin identified the potential importance of stratified analyses using propensity scores.[14] Our example shows both the usefulness of the test for interaction between propensity score and treatment effect and of describing the outcome of treatment across the range of propensity score values. If the propensity score analysis has worked as hoped we would expect to see a similar effect of 'treatment' across the range of propensity score values. If there are different effects for different propensity score values this should ring alarm bells, making it is a useful diagnostic.

Fourthly, and fundamentally, it is possible to identify important questions prospectively and conduct additional relevant randomised trials earlier, avoiding the need to rely on weaker methods. When clinicians need unbiased estimates of treatment effects among older people, women, or those with comorbidities, we should require answers to these questions from industry and publicly funded randomised trials.

Contributors: NF had the original idea for the paper, did the original statistical analyses described in the article, and wrote the first draft. JW aided in design. All authors revised the manuscript and approved the final version.

Competing interests: We have read and understood the BMJ policy on declaration of interests and have no relevant interests to declare.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Fisher RA. The design of experiments. 8th ed. Oliver and Boyd, 1966:21.
2 Avorn J. In defense of pharmacoepidemiology—embracing the yin and yang of drug research. *N Engl J Med* 2007;357:2219-21.
3 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41-55.
4 Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat Methods Med Res* 2012;21:273-93.
5 Smeeth L, Douglas I, Hall AJ, Hubbard R, Evans S. Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. *Br J Clin Pharmacol* 2008;67:99-109.
6 Pitt B, Annad FZ, Remme WJ, Cody R, Castaigne A, Perez A, et al. The effect of spironolactone on morbidity and mortality in patients with severe heart failure. *N Engl J Med* 1999;341:709-17.
7 Pitt B, Remme W, Zannad F, Neaton J, Martinez F, Roniker B, et al. Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *N Engl J Med* 2003;348:1309-21.
8 Zannad F, McMurray JJV, Krum H, van Veldhuisen DJ, Swedberg K, Shi H, et al. Eplerenone in patients with systolic heart failure and mild symptoms. *N Engl J Med* 2011;364:11-21.

9 Trends in the characteristics, treatment and prognosis of people with symptomatic heart failure and the use of aldosterone antagonists: analysis plan. www.ucl.ac.uk/priment/documents/Heart-Failure-Analysis-Plan.
10 Mangano DT, Tudor JC, Dietzel C. The risk associated with aprotinin in cardiac surgery. *N Engl J Med* 2006;354:353-65.
11 Mangano DT, Miao Y, Vuylsteke A, Tudor JC, Juneja R, Filipescu D, et al. Mortality associated with aprotinin during 5 years following coronary artery bypass graft surgery. *JAMA* 2007;297:471-9.
12 A comparison of rate control and rhythm control in patients with atrial fibrillation. *N Engl J Med* 2002;347:1825-33.
13 Saksena S, Slee A, Waldo AL, Freemantle N, Reynolds M, Rosenberg Y, et al. Cardiovascular outcomes in the AFFIRM trial: an assessment of individual antiarrhythmic drug therapies compared to rate control using propensity score matched analyses. *J Am Clin Cardiol* 2011;19:1975-85.
14 Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984;79:516-24.
15 Howell NJ, Senanayake EL, Freemantle N, Pagano D. Putting the record straight: aprotinin is safe and effective. Results from a mixed treatment meta-analysis of trials of aprotinin including the BART study. *J Thor Cardiovasc Surg* 2013;145:234-40.

Accepted: 11 September 2013

Cite this as: *BMJ* 2013;347:f6409

Figure

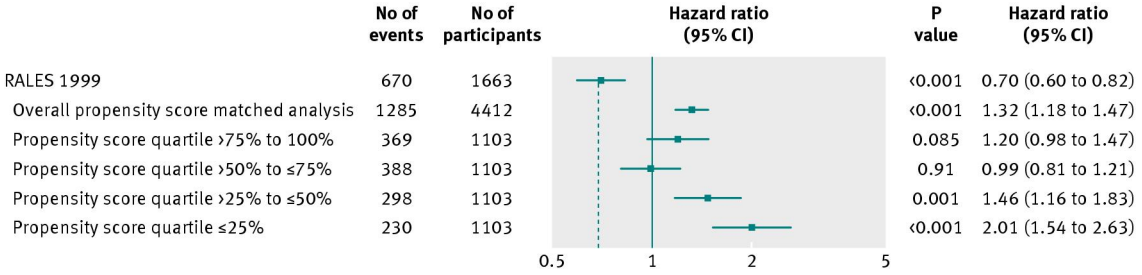

Fig 1 Effect of spironolactone on mortality by all cause in RALES and propensity score analysis. Results for propensity analysis given overall and by quartile