Methodological obstacles in causal inference: confounding, missing data, and measurement error

Penning de Vries, B.B.L.

6

Cautionary note: propensity score matching does not account for bias due to censoring

Bas B. L. Penning de Vries
Rolf H. H. Groenwold

Nephrology, Dialysis and Transplantation 2018; 33(6): 914–916

Abstract

This article gives a review of the limitations of propensity score matching as a tool for confounding control in the presence of censoring. Using an illustrative simulation study, we emphasize the importance of explicit adjustment for selective loss to follow-up and explain how this may be achieved.

In epidemiological research, valid causal inference is often hampered by confounding and selective loss to follow-up. Confounding is increasingly often addressed by means of propensity score (PS) matching. The analysis of a PS matched dataset closely resembles that of a randomised controlled trial (RCT); one expects that, on average, the distribution of covariates will be similar between treatment groups after propensity score matching or randomisation so that in the absence of other forms of bias systematic differences in outcomes between treatment groups can be attributed to treatment. Importantly, as is the case with RCTs (Groenwold et al., 2014). PS matching (or randomisation in the case of an RCT) typically does not account for selective loss to follow-up, and the confounder balance that was achieved through PS matching (or randomisation) may falsely reassure researchers and readers that the treatment groups under study were (and remained) comparable. The problem of selective loss to follow-up can, however, be potentially remedied by the same methods that have been proposed to address the problem in RCTs, namely inverse probability weighting, multiple imputation, or regression adjustment (Groenwold et al., 2014).

Two examples

In a study on the dose-response relationship between sulfonylurea derivatives (SU) and major adverse cardiovascular events in elderly patients with type 2 diabetes, patients were censored if they switched their treatment regimen (Abdelmoneim et al., 2016). Matching on a high-dimensional PS created treatment groups (high and low dose SU) that were very similar in terms baseline characteristics, including those reflecting disease severity, comedication use, and comorbidity state. Possibly, however, those who switched treatments at any point during follow-up represent a selective subset, for example because switching occurred more often among those who used more concomitant medication. Over time, this may have distorted the balance in comedication that was initially achieved through PS matching.

Another example is a study comparing outcomes between incremental and thrice-weekly initiation of haemodialysis (Park et al., 2016). Following PS matching, the groups were similar in terms of a number of baseline characteristics including age, sex, and primary renal disease. However, approximately half of the participants were lost to follow-up at 12 months. Again, this may have induced a selection bias if the loss to follow-up affected the treatment groups differentially.

An illustration of the problem

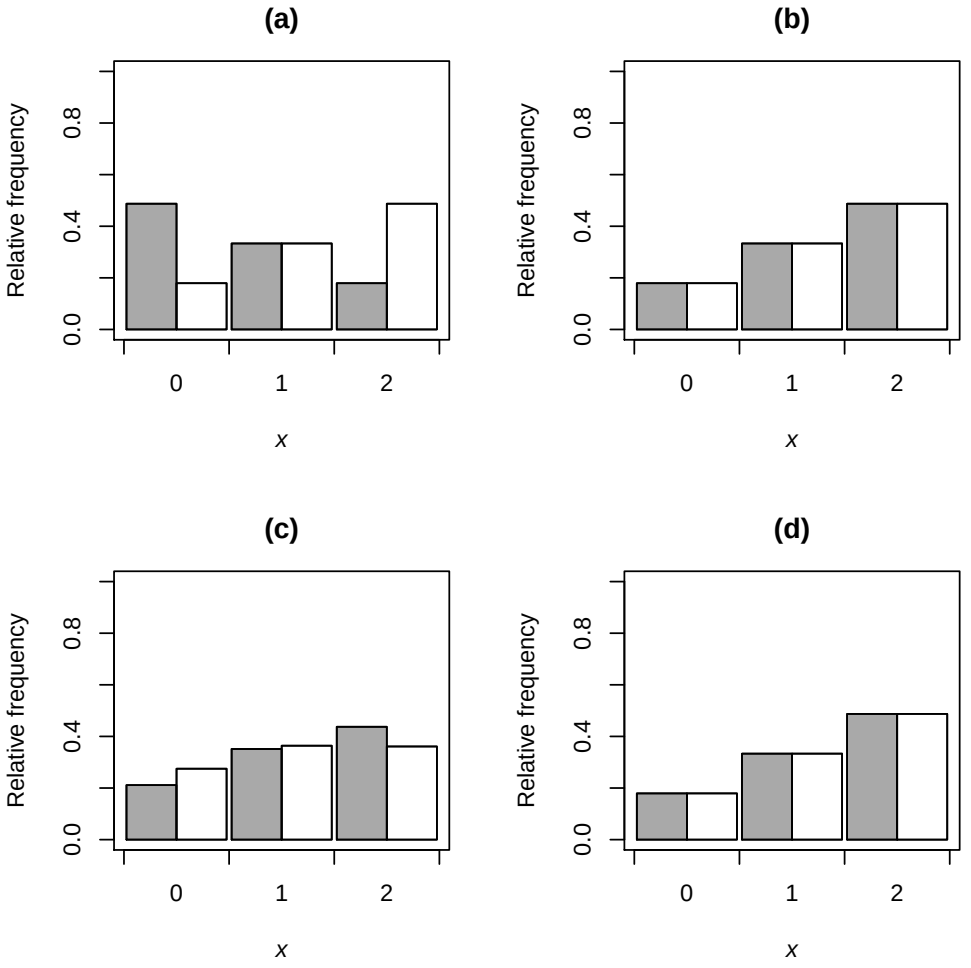
Through a small simulation study, we will illustrate the effect of ignoring selectively missing outcomes, whilst focusing on PS matching to control for confounding. Throughout, it is assumed that there is exchangeability for treatment and censoring, consistency, no model misspecification, and positivity, so that the observed covariates are sufficient to adjust for both confounding and selection bias due to loss to follow-up (Robins et al., 2000; Hernán et al., 2000; Cole and Hernán, 2008).

For this illustration, we consider a hypothetical setting representing an observational study of a binary treatment variable T, a binary outcome variable Y, and a trichotomous confounder X. The probability of a subject dropping out before their outcome could be assessed depends on both T and X. Data were generated for 10,000 subjects using the mechanism detailed in the Supplementary Material. The interest lies in the marginal odds ratio (OR) of 2 for the average treatment effect on the treated (ATT). However, in this observational setting, causal inference is hampered by confounding. This motivates the use of PS matching, which typically provides an estimate of the ATT (Williamson et al., 2012). Here, PSs were estimated by a logistic regression of T on X. We then matched treated to untreated subjects on the estimated PSs with replacement. As an alternative to PS matching to estimate the ATT, we also used inverse probability weighting, with weights of 1 and PS/(1−PS) for treated and untreated subjects, respectively. Treatment effects were estimated by applying a logistic regression to the matched or weighted pseudopopulations. We refer to these approaches as PS1 and IPW1, respectively. This procedure was repeated 1000 times. Bias was estimated on the log-odds ratio scale as the average deviation from the true log-odds ratio log 2.

The results in Table 6.1 show that both PS1 and IPW1 yielded substantial bias. The reason for this bias is apparent from Figure 6.1, which depicts the balance in the population before and after matching and/or weighting. Although PS1 and IPW1 are suited to balance confounders (Figure 6.1(a) and (b)), as subjects are lost to follow-up, the balance achieved through matching or weighting is not guaranteed to uphold in the dataset used for the analysis (Figure 6.1(c)). In fact, since the probability of dropping out depends on both T and X, conditioning on not being lost at follow-up (i.e. performing an analysis on those subjects for whom an outcome is observed) induces an association between X and T (Pearl, 2009; Hernán et al., 2004), thereby biasing the relation between T and Y through what is formally known as collider stratification bias.

To account for selective loss to follow-up, we applied Inverse-Probability-of-

Figure 6.1: Balance on the confounder X across treatment groups in a hypothetical setting



The untreated group is represented in grey; the treated group in white. Frequencies are relative to treatment (treated/untreated) group size; hence, equally sized bars indicate confounder balance. In the following, PS and PC denote the propensity score and the probability of censoring (being lost to follow-up) given T and X, respectively.

Panel (a) shows the balance in the original unweighted population. Reweighting observations using weights of 1 and $PS/(1-PS)$ for treated and untreated subjects, respectively, results in the balance shown in (b). The same result is obtained by matching treated subjects to untreated with similar PS. Removing observations with censored outcomes from this inverse probability weighted or PS matched dataset results in imbalance (c). The balance shown in (d) is obtained by weighting the original observations with $1/(1-PC)$ and $PS/[(1-PC)(1-PS)]$ for treated and untreated subjects, respectively, and conditioning on noncensored observations. The same result is obtained by reweighting the PS matched dataset by $1/(1-PC)$ for each subject.

Censoring-Weighting (IPCW) (Robins et al., 2000; Cole and Hernán, 2008). In this simple setting with only one point of follow-up, the IPCW weights reduce to the inverse probability of not being lost to follow-up (censored). Probabilities of censoring (PC) were estimated by logistic regression of C, a censoring indicator, on T and X applied to the original datasets. We then applied two additional estimators, PS2 and IPW2. In PS2, the matched sets obtained through PS1 were additionally weighted by $1/(1-PC)$ for each subject. In IPW2, the weights $1/(1-PC)$ and $PS/[(1-PS)(1-PC)]$ for the treated and untreated, respectively, were applied to the original datasets, and only subjects with observed outcomes were included in the analysis. Again, treatment effects were estimated by applying a logistic regression to the matched and/or weighted pseudopopulations.

The results in Table 6.1 show that both PS2 and IPW2 yielded estimates that on average were very close to the true effect. The reason is that PS2 and IPW2

Table 6.1: Performance of inverse probability weighting (IPW) and PS matching estimators

Estimator	Bias (95%CI)	OR
PS1	−0.134 (−0.139, −0.129)	1.749
IPW1	−0.135 (−0.139, −0.130)	1.748
PS2	0.002 (−0.003, 0.008)	2.004
IPW2	0.002 (−0.003, 0.007)	2.003

For definitions of PS1, IPW1, PS2, and IPW2, see text. Bias was estimated by the average deviation of the estimated log-odds ratios $\hat{\beta}$ from the true effect $\beta = \log 2$ across 1000 simulated samples. 95%CI $= \bar{\hat{\beta}} - \beta \pm 1.96\sqrt{(\hat{\sigma}^2/1000)}$, where $\hat{\sigma}^2$ denotes the empirical variance of $\hat{\beta}$. OR $= \exp \bar{\hat{\beta}}$ (True OR = 2).

restore the imbalance that resulted from conditioning on not being lost to follow-up by reweighting observations such that X and T are no longer associated, and X takes the distribution of the treated subjects (Figure 6.1(d)).

Covariate imbalance in the absence of censoring

It should be borne in mind that with two or more points of follow-up, covariate imbalance can develop even in the absence of censoring—specifically, that is, leaving the risk set for reasons other than sustaining the outcome of interest. Conditioning on past survival may induce an association between treatment and marginally independent covariates if past survival is a common effect of both (Hernán et al., 2004; Hernán, 2010; Aalen et al., 2015; Sjölander et al., 2016). If these covariates are also predictive of survival at a subsequent point of follow-up, this conditioning may therefore open a backdoor path, thereby inducing a selection bias. Thus, neither RCTs or PS matching or weighting analyses are guaranteed to be free of selection bias, because such selection occurs after baseline imbalances have been removed through randomisation, matching or weighting.

Conclusion

PS methods have gained increasing interest as means to adjust for confounding (Stürmer et al., 2006). However, as illustrated, PS matching does not account for bias due to censoring. In fact, the balance of confounders across treatment groups that was achieved by PS matching may be ruined by selective censoring. This problem can potentially be remedied by inverse probability of censoring weighting (as shown here), multiple imputation, or regression adjustment. It is important to be aware, however, that in contrast to PS matching and inverse probability weighting, the estimand of conventional multivariable regression analysis is not typically a marginal effect such as the ATT. Also, our simulations were done under the assumption that the censoring mechanism was independent of the outcome. Importantly, neither of the above methods is suited to solve the problem of censored data when the missingness depends on unobserved variables that are predictive of the outcome or on the outcome itself. It is only when the missingness can be explained by observed data, such as in our illustration, that such biases may be adequately addressed by one of the above methods. If loss to follow-up is a completely random process, the confounder balance that was achieved by PS matching is expected to be preserved and conventional analysis on those for whom an outcome was observed will still be appropriate.

References

Aalen, O. O., R. J. Cook, and K. Røysland (2015): "Does cox analysis of a randomized survival study yield a causal treatment effect?" *Lifetime data analysis*, 21, 579–593.

Abdelmoneim, A. S., D. T. Eurich, A. Senthilselvan, W. Qiu, and S. H. Simpson (2016): "Dose-response relationship between sulfonylureas and major adverse cardiovascular events in elderly patients with type 2 diabetes," *Pharmacoepidemiology and drug safety*.

Cole, S. R. and M. A. Hernán (2008): "Constructing inverse probability weights for marginal structural models," *American journal of epidemiology*, 168, 656–664.

Groenwold, R. H., K. G. Moons, and J. P. Vandenbroucke (2014): "Randomized trials with missing outcome data: how to analyze and what to report," *Canadian Medical Association Journal*, 186, 1153–1157.

Hernán, M. A. (2010): "The hazards of hazard ratios," *Epidemiology (Cambridge, Mass.)*, 21, 13.

Hernán, M. Á., B. Brumback, and J. M. Robins (2000): "Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men," *Epidemiology*, 11, 561–570.

Hernán, M. A., S. Hernández-Díaz, and J. M. Robins (2004): "A structural approach to selection bias," *Epidemiology*, 15, 615–625.

Park, J. I., J. T. Park, Y.-L. Kim, S.-W. Kang, C. W. Yang, N.-H. Kim, Y. K. Oh, C. S. Lim, Y. S. Kim, and J. P. Lee (2016): "Comparison of outcomes between the incremental and thrice-weekly initiation of hemodialysis: a propensity-matched study of a prospective cohort in korea," *Nephrology Dialysis Transplantation*, gfw332.

Pearl, J. (2009): *Causality: models, reasoning, and inference*, Cambridge university press.

Robins, J. M., M. A. Hernán, and B. Brumback (2000): "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 11, 550–560.

Sjölander, A., E. Dahlqwist, and J. Zetterqvist (2016): "A note on the noncollapsibility of rate differences and rate ratios," *Epidemiology*, 27, 356–359.

Stürmer, T., M. Joshi, R. J. Glynn, J. Avorn, K. J. Rothman, and S. Schneeweiss (2006): "A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods," *Journal of clinical epidemiology*, 59, 437–e1.

Williamson, E., R. Morley, A. Lucas, and J. Carpenter (2012): "Propensity scores: from naive enthusiasm to intuitive understanding," *Statistical methods in medical research*, 21, 273–293.

Supplementary Material

In our hypothetical setting, the mechanism for generating data is defined by sequentially sampling for each subject (independently) from the following distributions. Covariate X takes values 0, 1, and 2 only, each with probability 1/3. $T|X = x$ has the Bernoulli distribution with probability $\Pr(T = 1|x) = \text{expit}\{-1 + x\}$, were $\Pr(T = 1|x)$ is shorthand notation for $\Pr(T = 1|X = x)$. $C|x, t$ has the Bernoulli distribution with $\Pr(C = 1|x, t) = \text{expit}\{-1.5 + 0.5x + 2t\}$. Finally, $Y|x, t, c$ has the Bernoulli distribution with probability $\Pr(Y = 1|x, t, c) = \text{expit}\{-1 + x + 0.789t\}$. Potential outcomes $Y_{\check{t},\check{c}}$ under the combined treatment and censoring state $(\check{t}, \check{c})$ are distributed such that $\Pr(Y_{\check{t},\check{c}} = 1|x, t, c) = \Pr(Y = 1|x, t, c) = \text{expit}\{-1 + x + 0.789\check{t}\}$. By the law of total probability, $\Pr(Y_{\check{t},\check{c}} = 1|t) = \sum_{c=0}^{t}\sum_{x=0}^{2}\Pr(Y_{\check{t},\check{c}} = 1|x, t, c)\Pr(C = c|x, t)\Pr(X = x|t)$, where, by Bayes' theorem, $\Pr(X = x|t) = \Pr(T = t|x)\Pr(X = x)/\sum_{x=0}^{2}[\Pr(T = t|x)\Pr(X = x)]$. The interest lies in the marginal odds ratio θ for the treatment effect on the treated, if contrary to fact all subjects had remained uncensored; $\theta = \text{Odds}(\Pr(Y_{1,0} = 1|T = 1))/\text{Odds}(\Pr(Y_{0,0} = 1|T = 1))$, where $\text{Odds}(p) = p/(1 - p)$. It follows that $\theta \approx 2$.