Controlling Confounding Using Propensity Scores: Increasing Use, Persisting Questions, and New Venues

PD Dr. med. Til Stürmer

Brigham and Women's Hospital, Harvard Medical School

Propensity score analyses attempt to control for confounding in observational studies by adjusting for the likelihood that a given patient is treated. It has been proposed that such analyses achieve better control than multivariate outcome modeling in addressing confounding by indication, but there is little empirical evidence on this point. Based on a Medline search, we assessed the use of propensity scores over time and systematically evaluated those studies published in 2002. The use of propensity score analyses has increased sharply from a total of 5 papers before 1998 to 28 in 2002. Propensity scores varied considerably in their ability to predict treatment choice, with an area under the receiver operating characteristic curve ranging from a relatively poor value of 0.62 to the comparatively good value of 0.84. Two thirds of studies that presented results from both propensity score and 'traditional' multivariable outcome modeling showed very similar results. Use of propensity score methods has increased exponentially in non-experimental research, but there is little empirical evidence that these methods yield different or better estimates compared with 'traditional' multivariable outcome modeling. To study the comparative behavior of EPS and disease risk scores (DRS, combining several risk indicators into a single score), particularly with small study size, we then compared different scoring methods to control for confounding including propensity and disease risk scoring methods in evaluation of the effect of NSAID use on 1-year mortality from any cause in a cohort of 103,133 hospitalized elderly Medicaid beneficiaries. We chose this relation because it is subject to strong confounding, and the most plausible relation is at or near the null. From this cohort, we re-sampled 1,000 random subcohorts of 10,000, 1,000 and 500 people to assess the distribution of estimates. For each sample, we estimated the EPS and DRS using forward variable selection (alpha=0.3) and the 'traditional' multivariable outcome model using forward variable selection (alpha=0.2), additionally limiting the number of variables so as to have at least 8 outcomes per variable in the model. We used the estimated EPS to control for confounding by matching, by inverse probability of treatment weighting (IPTW), stratification, linear splines, and as a continuous variable in a proportional hazards outcome model. In the full cohort, the crude relative risk (RR) of dying for NSAID users was 0.68 (95% confidence interval: 0.66-0.71). The 'traditional' multivariate adjustment resulted in a RR of 0.80 (0.77-0.84). The RR closest to the most plausible truth of no protective effect of NSAID was achieved by IPTW (0.85;0.82-0.88). With decreasing cohort size, estimates remained further from the null, suggesting more residual confounding (despite an increasing c-statistic of the EPS predicting exposure), which was most pronounced for IPTW (for cohorts of N=500: RR = 0.72; 0.26 - 1.68). In this setting, the various ways to apply EPS and DRS behaved differently with smaller study size. Analytic techniques using EPS or DRS were not generally superior to 'traditional' multivariable outcome modeling. The c-statistic did not perform well in predicting the ability of EPS or IPTW in controlling confounding. Finally, I will present a new method to incorporate information on the joint distribution of unobserved confounders when assessing the sensitivity of effect estimates to unobserved confounding. This method is base on the combination of Propensity score methods and regression calibration.