Reproductive Toxicology 78 (2018) 75–80

Contents lists available at ScienceDirect

Reproductive Toxicology

journal homepage: www.elsevier.com/locate/reprotox

The impact of confounder selection in propensity scores when applied to prospective cohort studies in pregnancy

Ronghui Xu[a,b,*], Jue Hou[b], Christina D. Chambers[a,c]

[a] Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA, United States
[b] Department of Mathematics, University of California, San Diego, La Jolla, CA, USA
[c] Department of Pediatrics, University of California, San Diego, La Jolla, CA, United States

ARTICLE INFO

Article history:
Received 13 November 2017
Received in revised form 4 April 2018
Accepted 6 April 2018
Available online 7 April 2018

Keywords:
Inverse probability weighting
Regression adjustment
Standardization
Change in effects
p-value
Univariate association
Birth defects
Rare events
Average treatment effects

ABSTRACT

Our work was motivated by small cohort studies on the risk of birth defects in infants born to pregnant women exposed to medications. We controlled for confounding using propensity scores (PS). The extremely rare events setting renders the matching or stratification infeasible. In addition, the PS itself may be formed via different approaches to select confounders from a relatively long list of potential confounders. We carried out simulation experiments to compare different combinations of approaches: IPW or regression adjustment, with 1) including all potential confounders without selection, 2) selection based on univariate association between the candidate variable and the outcome, 3) selection based on change in effects (CIE). The simulation showed that IPW without selection leads to extremely large variances in the estimated odds ratio, which help to explain the empirical data analysis results that we had observed.

1. Introduction

Our work was motivated by work carried out at the Research Center for the Organization of Teratology Information Specialists (OTIS), which is a North American network of university or hospital based teratology services that counsel between 70,000 and 100,000 pregnant women every year. Among these women candidates for research studies are referred to the OTIS Research Center, and research subjects are also recruited through other methods including social media. Once the women are consented to a research study, the mothers and their babies are followed prospectively over time. Phone interviews are conducted throughout the duration of the pregnancy, along with pregnancy diaries recorded by the mother. An outcome telephone interview is conducted shortly after the pregnancy ends. If the pregnancy results in a live birth, a dysmorphology exam is performed within the first year of life and with further follow-ups at one year and possibly later dates.

The birth prevalence of major birth defects in the general population is about 3%, according to the Centers for Disease Control and Prevention (CDC) Metropolitan Atlanta Congenital Defects Program (MACDP), a population-based birth defects surveillance program and population-based references for secondary endpoints [1]. As pregnant women exposed to a specific medication or other substances in a given recruitment time period are often limited in number, sample sizes in these safety prospective cohort research studies are often limited to as few as 200 subjects in each exposure group, and are powered to detect an odds ratio (OR) of three or larger [2]. When there is no increased risk of birth defects, this often results in fewer than 10 events in each group.

In a typical research study of ours, the underlying scientific question of interest is whether a particular exposure causes major birth defects or other adverse outcomes, pregnant women are excluded from randomized clinical trials in the United States. The pregnancy outcomes research studies that we conduct are therefore observational in nature. As such, potential confounding is a main concern when drawing conclusions for a causal question of interest. In order to control for the potential confounding, a large number of maternal characteristic variables are typically collected. Given the limited sample sizes and in particular number of adverse events

* Corresponding author at: 9500 Gilman Drive, La Jolla, CA, 92093-0112, United States.
E-mail address: rxu@ucsd.edu (R. Xu).

https://doi.org/10.1016/j.reprotox.2018.04.003
0890-6238/

as explained above, it is difficult to directly adjust for so many covariates.

In the cases of observational studies with rare events such as birth defects, propensity score (PS) methods have been well established in the literature to count for potential confounding [3–5]. The propensity score is the conditional probability of receiving a treatment (i.e. exposure) given the observed covariates. It has the so-called balancing property that if strata or matched sets are formed that are homogeneous in the PS, even if they are heterogeneous in the covariates, the observed covariates will tend to balance [6]. The PS is typically computed by fitting a regression model of the exposure on the covariates; for example, for a binary exposure, a logistic regression model might be used. The analysis methods using the PS generally include matching, stratification, regression adjustment or inverse probability of treatment weighting (IPTW or IPW in general). IPW was initially prpoposed by Horvitz and Thompson [7] to weigh the observed responses in a survey by their inverse probabilities of inclusion (in the survey), and has since been widely used in missing data problems and causal inference [8,9]. The idea is a relatively straightforward one, since the IPW gives rise to a pseudo dataset that is a random sample of the population of interest. Tutorials on these methods using the PS are available in the literature; see for example D'Agostino, Jr. [10] and Austin [11].

As part of the OTIS Autoimmune Diseases in Pregnancy Project, medications used to treat a variety of autoimmune diseases in pregnancy are evaluated for safety using a prospective controlled cohort design comparing exposed to unexposed pregnancy outcomes. In a recent such study we had 319 pregnant women who were exposed to the medication under study and whose pregnancies ended in live birth, and 144 pregnant women who had the underlying diseases but were not exposed to the medication and whose pregnancies also ended in live birth. Out of these we had 30 major birth defects in the exposed group, and 5 major birth defects in the unexposed group. Due to the extremely rare events in our case, even matching or stratification using the PS becomes impractical, as they may lead to further deletion of observed events. IPW on the other hand, has become popular at least partly due to its ease of implementation, since most regression software allow weights as an option. Table 1 shows the results of analyses using either regression adjustment or IPW with stabilized weights [12–14] (that were further truncated to be between 0.1 and 10), with PS formed by change-in-estimate (CIE; see below) to confirm actual confounders or by simply including all potential confounders collected in the study without any selection or confirmation. The list of all potential confounders is provided in the Supplemental Materials. When including all potential confounders the sample size was slightly reduced due to missing values, leading to slightly different crude (i.e. unadjusted) odds ratio (OR) between exposure to the medication and the outcome of major birth defects. It is clear from the table that the IPW approach using all potential confounders gave an OR of 6.45, which was very different from the other estimated OR's.

Our main concern here is to what extent we should perform variable selection in computing the PS. In practice we don't know if an observed variable is truly a confounder [15], and different methods have been used to assess confounding. Two common approaches in practice are: 1) change-in-estimate (CIE) [16], which indirectly assesses the association of the candidate variable with both the exposure and the outcome, since a confounder should be a common cause of both; and 2) significance testing of the association between the candidate variable and the outcome only, which was recommended by Rubin [17] in order to reduce the variance of the estimated exposure effect. A third approach is to include all potential confounders. While it has been shown that variables that are only weakly associated with the outcome should not be included in the PS for small studies [18], this does not appear to be widely known and confusion persists in practice [19].

Table 1
Estimated Odds Ratio (95% CI) of Birth Defects with Different Approaches Using Propensity Scores.

	Crude	Reg. Adjustment	IPW
Selection by CIE	2.89 (1.10, 7.60)	2.74 (1.31, 8.71)	3.38 (1.27, 9.00)
All Potential Confounders	2.87 (1.09, 7.58)	2.94 (1.36, 9.98)	6.45 (2.26, 18.36)

In the following we describe the causal effects of exposure that we are interested in estimating, and carry out simulation experiments to study different approaches to estimate them. Among these approaches we will focus on regression adjustment and IPW using the PS.

2. Methods

Here we restrict our attention to a binary outcome, and the effect measure commonly used in practice is the OR. As logistic regression is commonly used and will be used to generate data here, we briefly discuss the non-collapsibility of logistic regression [15]. This can be briefly summarized as the discrepancy between the 'population averaged' effect and the 'conditional' effect under the logistic regression model given other covariates. Let $A = 1$ denote the exposed group, and 0 the unexposed group. The logistic regression model for the binary outcome Y is

$$P(Y = 1) = expit\left(\alpha_0 + \alpha_A A + \beta' X\right), \quad (1)$$

where $expit(x) = e^x/(1 + e^x)$ and X are the additional covariates. The coefficient α_A in the data-generating model, i.e. the conditional exposure effect given X, is sometimes used as the 'true' effect in simulation studies for assessing bias and estimation errors in general [18,20]. However, we note that it is not the probability limit to which the IPW estimator converges. In the Supplement Materials we show that IPW estimator converges to the logarithm of the marginal odds ratio between Y and A; this is also referred to as the average treatment effect (ATE) in the literature. This quantity does not generally have closed-form formula based on model (1) for a given distribution of X, but can be approximated using a very large Monte Carlo sample.

As an alternative to using the IPW to estimate the ATE, we can also use what is called standardization in the regression adjustment approach. That is, after fitting the outcomes regression model by including the exposure group indicator as well as the logit of the PS as regressors, we use the estimated regression coefficients to predict every subject's counterfactual probability of outcome under exposure by setting their $A = 1$; averaging over these counterfactual probabilities gives the marginal probability of outcome under exposure. We then similarly predict every subject's counterfactual probability of outcome under no exposure by setting their $A = 0$; averaging over these counterfactual probabilities gives the marginal probability of outcome under no exposure. These marginal probabilities are then used to calculate the marginal odd ratio of exposure.

For each simulation scenario below, we will compare the following estimates of the log odds ratio of exposure on outcome: crude, ignoring any covariate information; regression adjustment by including the logit of the PS as a linear term; IPW using PS; and fitting the multivariable logistic regression model with the true confounders but without the unobserved variables (see below). To form the PS, we consider four different ways of selecting confounders: 1) oracle, i.e. using the true confounders; 2) CIE, using at least 10% change as criterion in the estimated OR when adjusting for the potential confounder as compared to the crude OR; 3) sig-