Journal of Clinical Epidemiology 80 (2016) 97−106

Journal of Clinical Epidemiology

Propensity score model overfitting led to inflated variance of estimated odds ratios

Tibor Schuster[a,b,c,d,*], Wilfrid Kouokam Lowe[a,b,e], Robert W. Platt[b,f]

[a]*Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, 3755 Chemin de la Côte-Sainte-Catherine, Montréal, Québec H3T 1E2, Canada*

[b]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Purvis Hall, 1020 Pine Avenue West, Montréal, Québec H3A 1A2, Canada*

[c]*Clinical Epidemiology and Biostatistics Unit and the Melbourne Children's Trial Centre, Murdoch Childrens Research Institute, Royal Children's Hospital, 50 Flemington Road, Parkville, Victoria 3052, Australia*

[d]*Department of Paediatrics, University of Melbourne, Melbourne, Victoria 3010, Australia*

[e]*UFR de Mathématique et d'Informatique, Université de Strasbourg, 7 Rue René Descartes, 67084 Strasbourg, France*

[f]*Department of Pediatrics, McGill University, Montreal Children's Hospital, 1001 Décarie Boulevard, Montreal, Québec H4A 3J1, Canada*

Accepted 11 May 2016; Published online 14 September 2016

Abstract

Objective: Simulation studies suggest that the ratio of the number of events to the number of estimated parameters in a logistic regression model should be not less than 10 or 20 to 1 to achieve reliable effect estimates. Applications of propensity score approaches for confounding control in practice, however, do often not consider these recommendations.

Study Design and Setting: We conducted extensive Monte Carlo and plasmode simulation studies to investigate the impact of propensity score model overfitting on the performance in estimating conditional and marginal odds ratios using different established propensity score inference approaches. We assessed estimate accuracy and precision as well as associated type I error and type II error rates in testing the null hypothesis of no exposure effect.

Results: For all inference approaches considered, our simulation study revealed considerably inflated standard errors of effect estimates when using overfitted propensity score models. Overfitting did not considerably affect type I error rates for most inference approaches. However, because of residual confounding, estimation performance and type I error probabilities were unsatisfactory when using propensity score quintile adjustment.

Conclusion: Overfitting of propensity score models should be avoided to obtain reliable estimates of treatment or exposure effects in individual studies.

Keywords: Propensity score; Logistic regression; Overfitting; Confounder adjustment; Odds ratio; Inverse probability weighting

1. Introduction

Observational studies are frequently used to estimate treatment or exposure effects in settings where the assignment of subjects into intervention or exposure groups is not under control of the study investigator. A major shortcoming of such studies is that treatment preference or the status of exposure is often linked to individual characteristics that are not independent of the outcome of interest. Therefore, comparison groups may differ in their covariate distributions in ways that will confound the results regarding estimated treatment or exposure effects on the outcome.

Propensity scores can be used to aggregate information about the predictive role of covariates on treatment assignment or exposure status. Formally, the propensity score is

Funding: T.S. and W.K.L. were supported by the Canadian Network for Observational Drug Effect Studies (CNODES). CNODES is funded by a grant from Health Canada, the Drug Safety and Effectiveness Network (DSEN), and the Canadian Institutes of Health Research (CIHR) grant number 111845-1. T.S. was also supported by funding from the Royal Children's Hospital Foundation to the Melbourne Children's Trial Centre. Research at the Murdoch Childrens Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program. R.W.P. is supported in part by a National Scholar (Chercheur-national) of the Fonds de Recherche du Quebec−Sante (FQR-S) and is a member of the Research Institute of the McGill University Health Center, which is supported by core funds from FQR-S.

* Corresponding author. Tel.: +61-3-9936-6097; fax: +61-3-9348-1391.

E-mail address: Tibor.Schuster@mcri.edu.au (T. Schuster).

http://dx.doi.org/10.1016/j.jclinepi.2016.05.017

What is new?

Key findings

- Overfitting of propensity score models leads to inflation of the variance of effect estimates when applying established conditional and marginal inference methods that use propensity scores for confounder adjustment.

What this adds to what was known?

- Consequently, estimate uncertainty obtained in an individual study can annul alleged unbiasedness due to confounding control if the number of exposed or unexposed individuals per propensity score predictor variable is low.
- Conventional propensity score quintile adjustment is less effective in confounding control than conditioning on propensity score spline functions or using inverse probability of treatment (exposure) weighting.

What is the implication and what should change now?

- We recommend that specification of propensity score models should acknowledge widely accepted guidelines for regression model building to avoid overfitting.
- We discourage the use of propensity score quintile adjustment in favor of modeling propensity score spline functions or using inverse probability of treatment (exposure) weighting.

the probability of receiving treatment (or experiencing a certain exposure status) given individual covariate realizations [1]. There are different ways to use propensity scores to address confounding such as matching based on the propensity score, stratification according to propensity score intervals, ordinary propensity score adjustment in the context of a multivariable binary logistic regression analysis, and performing weighted effect estimation (inverse probability of treatment weighting) in the framework of marginal structural models [2,3].

Because propensity score modeling is undertaken to aggregate multivariate covariate information into a single variable, propensity methods are particularly popular when estimating treatment or exposure effects on rare outcomes using data sets with a large number of potential confounding variables. Binary logistic regression is the most common model used to estimate propensity scores. Previous simulation studies have shown that the number of events relative to the number of parameters in the logistic model should exceed a ratio between 10 or 20 to 1 to avoid inflated standard errors of the parameter estimates [4–6]. Further simulation-based investigations have demonstrated that this rule may be relaxed in sensitivity analyses to demonstrate adequate control of confounding [7].

Although there is an ongoing debate and controversy in the literature about correct propensity score model specification, only limited research has been undertaken yet to systematically investigate the role of overfitting logistic propensity score models that are incorporated in different conditional and marginal inference approaches [8–13]. Available simulation studies on the number of variables included in the propensity score did not directly consider the ratio of number of exposed or treated individuals to propensity score predictor variables and were based on real data without knowledge of the true effect of treatment on the outcome [14].

In fact, there is a wide-spread perception that the propensity score is meant to be only descriptive for the data in hand but not to be generalizable to other data sets [15]. We investigate within this article whether inaccurate estimation of the propensity score due to model overfitting leads to considerable bias or inflated variance of estimated effect parameters.

The article is structured as follows: in Sections 2 and 3, we describe the designs of comprehensive Monte Carlo and plasmode simulation studies that investigate to which extent overfitting of propensity score models leads to systematically and randomly erroneous effect estimates. In Section 4, we report the resulting bias, root mean square error, as well as type I and type II error rates in testing the null hypothesis of no treatment effect. Section 5 closes with the discussion of the results and conclusions.

2. Monte Carlo simulation setup

2.1. General data scenario and inference methods to be compared

We consider the scenario of a point-exposure study investigating the effect of a binary treatment E on a dichotomous outcome Y. Within this study, a binary logistic regression model (the propensity score model) is used to estimate every study individual's probability of receiving treatment given the realizations of a prespecified set of covariates $X_1, \ldots, X_k$. The respective propensity score is then used in different ways to account for potential confounding when estimating the conditional or marginal odds ratio as effect parameter. In particular, we consider the following effect estimation approaches within our study: (A) multivariable logistic regression for the binary outcome to estimate the conditional treatment effect (log odds ratio) under adjustment for the entire set of covariates, (B) multivariable logistic regression for the binary outcome to estimate the treatment effect conditioning on binary variables that indicate an individual's membership to one of the quintile-based partitions of the estimated propensity score