Published in final edited form as:

Ann Epidemiol. 2020 January ; 41: 56–59. doi:10.1016/j.annepidem.2019.12.006.

Visualization tool of variable selection in bias–variance tradeoff for inverse probability weights

Ya-Hui Yu, PhD^{a,b}, Kristian B. Fillion, PhD^{a,b,c}, Lisa M. Bodnar, PhD^{d,e,f}, Maria M. Brooks, PhD^d, Robert W. Platt, PhD^{a,b,g,h}, Katherine P. Himes, MD^{e,f}, Ashley I. Naimi, PhD^{d,*}

^aDepartment of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada

^bCentre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, Montreal

^cDepartment of Medicine, McGill University, Montreal, QC, Canada

^dDepartment of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA

^eDepartment of Obstetrics, Gynecology, and Reproductive Sciences, School of Medicine, University of Pittsburgh, Pittsburgh, PA

^fMagee-Womens Research Institute, Pittsburgh, PA

^gMcGill University Health Center Research Institute, Montreal, QC, Canada

^hDepartment of Pediatrics, McGill University, Montreal, QC, Canada

Abstract

Purpose: Inversed probability weighted (IPW) estimators are commonly used to adjust for time-fixed or time-varying confounders. However, in high-dimensional settings, including all identified confounders may result in unstable weights leading to higher variance. We aimed to develop a visualization tool demonstrating the impact of each confounder on the bias and variance of IPW estimates, as well as the propensity score overlap.

Methods: A SAS macro was developed for this visualization tool and we demonstrate how this tool can be used to identify potentially problematic confounders of the association of statin use after myocardial infarction on one-year mortality in a plasmode simulation study using a cohort of 39,792 patients from the UK (1998–2012).

Results: Through the tool's output, we can identify problematic confounders (two instrumental variables) and important confounders by comparing the estimated psuedo MSE with that from the fully adjusted model and propensity score overlap plot.

*Corresponding author. Department of Epidemiology, University of Pittsburgh, 130 DeSoto St, 5131 Parran Hall, Pittsburgh, PA 15261. Tel.: +1-412-624-3174. ashley.naimi@pitt.edu (A.I. Naimi).

Conflict of interest: None declared.

Supplementary Data

Supplementary data associated with this article can be found in the online version at <https://doi.org/10.1016/j.annepidem.2019.12.006>.

Conclusion: Our results suggest that the analytic impact of all confounders should be considered carefully when fitting IPW estimators.

Keywords

Inverse probability weighting; Bias-variance tradeoff; Visualization; High-dimensional covariates

Introduction

Proper control of confounding is essential for estimating valid exposure effects in observational studies. Directed acyclic graphs (DAGs) [1] have generally been preferred over statistical modeling strategies (such as stepwise regression) for confounder identification, as the latter are known to yield biased exposure effect estimates and anticonservative measures of uncertainty [2,3]. DAG-based confounder selection requires content expertise and background knowledge to avoid adjusting for variables that may introduce bias and to appropriately adjust for a minimally sufficient set of variables that remove bias. However, issues may arise when the minimally sufficient set of adjustment variables is large relative to the sample size. In particular, adjusting for a larger number of confounding variables generally decreases the estimation bias but incurs a variance penalty such that the overall mean squared error of the estimator is higher (worse) than if the confounder were excluded [4]. This well-known bias-variance tradeoff can have an important impact on the quality of inferences from any study [5].

Inversed probability weighted (IPW) is one of the approaches often used in confounding adjustment, particularly in the presence of time-varying confounding affected by prior exposures [5]. Provided all confounders are properly accounted for, weighting creates a “pseudo-population” in which no confounding is present [6]. Weights for each study subject are constructed by taking the inverse probability of receiving their observed exposure conditional on confounders. However, when the probability of being exposed or unexposed is close to zero or one, IPW estimators are known to suffer from problems due to highly variable weights [7]. These problems are worsened in the presence of a high-dimensional confounder set.

To date, a limited number of diagnostic tools are available to signal problems with IPW estimators. These include simply taking the mean of the stabilized IPWs [8], evaluating the maximum IPW value, and visualizing the overlap in the propensity score between exposure groups. All such tools have relied on evaluating some function of the propensity score [9]. Here, we propose an algorithm that extends this evaluation to the estimand of interest, such as the average treatment effect or the effect of treatment on the treated. We developed a visualization tool that will allow researchers to evaluate the bias–variance tradeoff incurred by including or excluding a confounder from the propensity score model when estimating treatment effects using IPW.

Methods

SAS macro: %bias_var

Using the SAS programming language (SAS Institute, Cary, NC), we developed a *%bias_var* macro, which is designed for scenarios where there is a binary exposure and outcome at a single time point (Web Appendix). Propensity score models for binary exposures were constructed using logistic regression to predict the probability of being exposed after adjusting for covariates. For those who were exposed, the denominator of the IPW is the predicted probability of being exposed; for those who were not exposed, we used one minus the probability of being exposed in the denominator. To stabilize the weights, the unconditional probability of being exposed/unexposed was used as the numerator for those who were exposed/unexposed.

Following Greenland et al.[3], we computed a pseudo mean squared error (MSE) as a summary statistic for bias and variance of the IPW estimate. We assumed that the estimate from the fully adjusted model (as determined via our DAG) is the best proxy for the unbiased (true) estimate. We calculated bias as the deviance between this “unbiased” estimate with the estimate from model excluding a certain confounder (reduced model); standard error of this estimate from this reduced model was used to compute the variance.

$$\text{PseudoMSE} = (\beta_{\text{reduced}} - \beta_{\text{full}})^2 + (SE_{\text{reduced}})^2 \quad (1)$$

In addition, we also explored the estimates from two-sided IPW truncation at first and 99th or fifth and 95th percentiles of the distribution of the propensity score [8] as a comparison with the approach of removing confounders.

The required input parameters for the macro are as follows: data set, list of categorical confounders, list of continuous confounders, binary exposure variable, binary outcome variable, and unique identifier for each observation. This macro generates two plots: the first plot displays the risk ratios with 95% confidence interval from models that excluded the variable shown on the y-axis. The second plot shows the distribution (min, 25th, 50th, and 75th percentile, mean and max) of propensity score of exposed and nonexposed groups from the corresponding models. These plots are both sorted by ascending values of pseudo MSEs of the outcome model. The flow chart of each step in this macro is described in Web Figure 1.

Empirical setting

We applied our SAS macro *%bias_var* to an empirical study using data from a population-based cohort of the Clinical Practice Research Datalink and Hospital Episode Statistics to investigate the association of statin use after myocardial infarction (MI) on the risk of one-year mortality. Our analytic cohort consisted of patients aged 18 or older with a recorded MI between April 1st, 1998, and March 31st, 2012 (32,792 patients). Details on the process of constructing our analytic cohort were described previously [10]. Briefly, the cohort entry date was 30 days after the dates of either hospital discharge or of MI diagnosis. Patients treated with statin were defined as those who received statin within 30 days after the index

MI. The outcome of interest was all-cause mortality identified by any death recorded in the database during the one-year follow-up period. This study was approved by the Independent Scientific Advisory Committee of the Clinical Practice Research Datalink (protocol number: 14_018A4) and the Research Ethics Board of the Jewish General Hospital (Montreal, Quebec).

We used causal diagrams to identify all potential confounders affecting both receiving statin and death, including demographic characteristics (age, sex), year of cohort entry, lifestyle variables (smoking, alcohol use, obesity), comorbidities (atrial fibrillation, coronary artery disease, cerebrovascular disease, congestive heart failure, chronic obstructive pulmonary disease, diabetes mellitus, hypertension, hypercholesterolemia, peripheral vascular disease, previous coronary revascularization, previous stroke, previous MI), history of medication use (aspirin, angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, beta-blockers, calcium-channel blockers, diuretics, fibrates, and nonsteroidal anti-inflammatory drugs), and proxies of overall health (numbers of medications prescribed, numbers of hospitalization of previous year).

To evaluate impact of including instrumental variables in the propensity score model used to generate weights on the IPW estimator, we adopted a plasmode simulation approach to generate a new data set based on the aforementioned cohort. A plasmode simulation approach generates data sets with known data generating process of exposure and outcome while keeping the association among other covariates in the original data set [11]. It is a simulation approach often used to evaluate methods in the high-dimensional covariates scenario. First, we generated four binary variables (X1, X2, X3, and X4) from Bernoulli distribution with a probability of 0.5. These four variables were treated as two instrumental variables (X1, X2) and two confounders independent from other potential confounders (X3, X4) in later data generation steps. Second, a treatment variable was generated by a logistic regression including all the potential confounders and newly generated four covariates. The coefficients of potential confounders were obtained from the estimated coefficients from original data sets while the coefficients for X1 to X4 were 2, 1, 0.25, and 0.25, respectively. Finally, we used the aforementioned generated treatment variable and confounders to generate the binary outcome. Similarly, the coefficients for treatment were set to -0.7, and potential confounders were obtained from the original data set while the coefficients for X3 and X4 were 1 and 0.5, respectively. By contracting two marginal potential outcomes of two treatment groups, the treatment effect of this study was 0.62.

For simplicity, we randomly selected one of the simulated data sets for demonstration. Unlike a typical simulation study, we did not use the true treatment effect to calculate bias for our tool, instead we used the estimate from the model adjusting all covariates as the “true value”. However, if the propensity score model was correctly specified, this value should be close to the simulated effect. In total, 27 categorical variables along with the four newly generated variables were considered as confounders of the association of statin use after MI and one-year mortality. The macro % bias_var was applied to this simulation data set to identify the impact of each covariate on the bias–variance tradeoff of the estimate of interest.

Results

Plots of the adjusted risk ratios estimated after excluding each confounder sorted by pseudo MSE are provided in Figure 1A. In our example, the strong instrument (X1) was identified through our tool as ranked first place in the plot. This means excluding X1 generates the smallest pseudo MSE compared with all other models and improves the propensity score overlap. The exclusion of X1 produced a relative risk for all-cause mortality for statin versus no statin of 0.61 (95% CI: 0.56, 0.66) compared with a risk ratio from the fully adjusted model of 0.59 (95% CI: 0.53, 0.66) (Fig. 1B). When the two known instrumental variables (X1, X2) were excluded from the propensity score, the tool demonstrated that all models had smaller pseudo MSEs and better propensity score overlap (Fig. 2). Through the tool's output, we can identify important confounders by comparing the estimated pseudo MSE with that from the fully adjusted model. For example, removing age from propensity score model greatly increased the pseudo MSE, which implies that it is an important confounder that should be kept in the model. X3 and X4 are the confounders that we introduced in the data generation process, and they are not correlated with other covariates; therefore, removing them from the propensity score model also increased the pseudo MSE of the estimate.

Discussion

We developed a tool to visualize the impact of each confounder on characteristics of the point estimate of interest and applied it to estimating the IPW association between statin use and one-year all-cause mortality in a plasmade study using a cohort of 37,792 post-MI patients from the UK. Our results demonstrated the advantages of adopting this visualization tool for identifying highly influential variables based on the bias and variance of the point estimate and the overlap in the propensity score of the exposure of interest.

This user-friendly macro can serve as the first step to visually diagnose the impact of each confounder on the point estimates as well as propensity score distributions. When presented with a confounder causing positivity violation, a handful of options exist to manage its impact. One option is to remove the confounder from the model. Provided the actual confounding impact is negligible, this is the simplest approach. For a true confounder, however, one may use doubly robust collaborative targeted minimum loss-based estimation (cTMLE) [12]. With cTMLE, one may adjust for the confounder in the outcome model but exclude it from the propensity score model with sufficient confounder set. In doing so, confounding is adjusted for without suboptimal performance caused by nonoverlapping propensity scores. There is a package available in R to apply cTMLE [13].

Of note, for the purpose of comparing bias-variance tradeoff between models, we used the standard errors computed from the models for pseudo MSE calculation. To obtain more accurate calculation of variance, bootstrapping or cross-validation approaches can serve as alternatives [5].

Estimate from the fully adjusted model will be approximately unbiased only under the assumption that the model, measured confounders, and sample size are sufficient to estimate

the true effect [3]. We used this value as the “unbiased” estimate in our tool because it is the best available proxy. The propensity scores aim to create balanced groups between exposed and unexposed groups in terms of covariates. Our assessment of the impact of each confounder on the point estimate might conflict with the purpose of the propensity score as a “design” phase tool [14]. However, our tool addressed the issue that selection of variables for propensity scores plays an important role in obtaining unbiased and efficient estimates. It is especially useful for identifying extreme violations of the positivity assumption.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Filion is supported by a salary support award from the Fonds de recherche du Québec – santé (FRQS; Quebec Foundation for Health Research) and a William Dawson Scholar award from McGill University.

References

- [1]. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol* 2008;8:70. [PubMed: 18973665]
- [2]. Greenland S, Pearce N. Statistical Foundations for Model-Based Adjustments. *Annu Rev Public Health* 2015;36:89–108. [PubMed: 25785886]
- [3]. Greenland S, Daniel R, Pearce N. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *Int J Epidemiol* 2016;45: 565–75. [PubMed: 27097747]
- [4]. Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ* 2016;352:i1981. [PubMed: 27121591]
- [5]. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. 2nd ed. New York (NY): Springer-Verlag; 2009 <https://www.springer.com/gp/book/9780387848570>. [Accessed 28 March 2019].
- [6]. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–60. [PubMed: 10955408]
- [7]. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol* 2010;171:674–7. discussion 678–681. [PubMed: 20139125]
- [8]. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008;168:656–64. [PubMed: 18682488]
- [9]. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med* 2015;34: 3661–79. [PubMed: 26238958]
- [10]. Pang M, Schuster T, Filion KB, Schnitzer ME, Eberg M, Platt RW. Effect Estimation in Point-Exposure Studies with Binary Outcomes and High-Dimensional Covariate Data e A Comparison of Targeted Maximum Likelihood Estimation and Inverse Probability of Treatment Weighting. *Int J Biostat* 2016;12(2).
- [11]. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal* 2014;72:219–26. [PubMed: 24587587]
- [12]. van der Laan MJ, Gruber S. Collaborative double robust targeted maximum likelihood estimation. *Int J Biostat* 2010;6(1).
- [13]. Gruber S, van der Laan MJ. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat* 2010;6(1).
- [14]. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med* 2007;26: 20–36. [PubMed: 17072897]

Fig. 1.

Risk ratios for one-year all-cause mortality among receiving statin versus without receiving statin by models with full model of adjusting all the confounders. (A) Impact of excluding variables on bias and variance. Each row represented the risk ratio of stillbirth from applying the propensity score model adjusting for all the identified covariates except the one denoted on the left-hand side. RMSE = root of mean squared errors. Truncated 5% and truncated 1% represented the risk ratio of stillbirth with the truncated weights at first or fifth percentiles which are the default in our tool. (B) Impact of excluding variables on propensity score overlap. Propensity score distributions including minimum, first quartile, median, mean (denoted as empty dot), third quartile and maximum by exposure groups.

Fig. 2.

Risk ratios for one-year all-cause mortality among receiving statin versus without receiving statin by models with full model of adjusting all the confounders excluding instrumental variables (X1, X2). (A) Impact of excluding variables on bias and variance. Each row represented the risk ratio of stillbirth from applying the propensity score model adjusting for all the identified covariates except the one denoted on the left-hand side. RMSE = root of mean squared errors. Truncated 5% and truncated 1% represented the risk ratio of stillbirth with the truncated weights at first or fifth percentiles which are the default in our tool. (B) Impact of excluding variables on propensity score overlap. Propensity score distributions including minimum, first quartile, median, mean (denoted as empty dot), third quartile and maximum by exposure groups.