The International Journal of Biostatistics

Volume 6, Issue 2

2010

Article 15

CAUSAL INFERENCE

Impact of Outcome Model Misspecification on Regression and Doubly-Robust Inverse Probability Weighting to Estimate Causal Effect

Geneviève Lefebvre, *Université du Québec à Montréal*
Paul Gustafson, *University of British Columbia*

Recommended Citation:

Lefebvre, Geneviève and Gustafson, Paul (2010) "Impact of Outcome Model Misspecification on Regression and Doubly-Robust Inverse Probability Weighting to Estimate Causal Effect," *The International Journal of Biostatistics*: Vol. 6: Iss. 2, Article 15.

DOI: 10.2202/1557-4679.1207

Impact of Outcome Model Misspecification on Regression and Doubly-Robust Inverse Probability Weighting to Estimate Causal Effect

Geneviève Lefebvre and Paul Gustafson

Abstract

Estimating treatment effects with observational data requires adjustment for confounding at the analysis stage. This is typically done by including the measured confounders along with the treatment covariate into a regression model for the outcome. Alternatively, it is also possible to adjust for confounding by taking into account the propensity of an individual to receive treatment, with inverse probability weighting (IPW). In the class of IPW estimators, the so-called doubly-robust estimator also requires the specification of the outcome regression model, in addition to the propensity model. The aim of this paper is to investigate the impact of misspecification of the outcome model on the performances of the usual regression and doubly-robust IPW estimators for estimating treatment effects. We examine the performances of the estimators across the parameter space for different scenarios of model misspecification using large-sample theory. We find that for small-to-moderate sample sizes, the regression estimator compares favorably to the IPW doubly-robust estimator. Finally we argue, both conceptually and on the basis of our results, that treatment-confounder interactions should be included in the outcome regression model.

KEYWORDS: causal inference, regression, inverse probability weighting, double-robustness, model misspecification, effect-modification

Author Notes: This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Canadian Institutes of Health Research (CIHR).

1 Introduction

Controlling for confounding in observational studies has been traditionally accomplished by including the measured confounders along with the treatment covariate into a regression model for the outcome. Alternatively, it is also possible to adjust for confounding by taking into account the propensity of an individual to receive treatment, such as with inverse probability weighting (IPW) or propensity scores (PS) (Rosenbaum and Rubin, 1983; Rosenbaum, 1987; Hirano and Imbens, 2001; D’Agostino Jr., 1998). Although controlling for confounding using IPW or PS is not a new idea, it has become an increasingly popular alternative over the last years in medical research (Stürmer *et al.*, 2006a). There has also been an interest in considering so-called “doubly-robust” propensity score based estimators (Robins *et al.*, 1994; Tan, 2006), which are augmented to include some information about the outcome and treatment-confounders relationship. As for the regression estimator, these estimators also require that the analyst specifies an outcome regression model.

While some recent simulation studies have compared doubly-robust propensity score estimators and regression (Lunceford and Davidian, 2004; Kang and Schafer, 2007), we believe it worthy to further shed light on these estimators for causal treatment effect estimation. With this objective in mind, the aim of the paper is to investigate to what extent the misspecification of the outcome model impacts on the regression and doubly-robust IPW estimators. In particular, we give attention to a special form of model misspecification for the outcome regression model. At least from a conceptual point-of-view, we argue that this model should include interaction terms between treatment and confounders. However, outcome regression models that allow covariates to be simultaneously confounders and effect-modifiers are not standard in practice. Finally, the use of large-sample theory facilitates the examination of the performance of these estimators across the parameter space.

2 Background

2.1 Causal Effects Estimation

Rubin’s counterfactual framework (Holland, 1986) is convenient to formulate the problem and define the causal quantity of interest. For simplicity, we restrict our attention to the estimation of the effect of a non time-varying treatment on the mean of a continuous outcome. Available data are (Y, X, C) for a random sample of size n from the population, where Y is the continuous outcome, X is the binary treatment indicator and C are the measured

confounders. To each individual, we also associate the so-called *counterfactual* outcomes (Y_1, Y_0) , where Y_x , $x = 0, 1$, is the value of the outcome that would be observed, if, possibly contrary to the fact, the individual were allocated to treatment $X = x$. The goal is to estimate the average difference in outcome that is due to the treatment, $\Delta = \mu_1 - \mu_0 = E[Y_1 - Y_0]$. One problem with this formulation is that only Y_1 or Y_0 is observed for a given individual, assuming that $Y = Y_x$ when $X = x$ (consistency assumption). Therefore the estimation of Δ is a non-identifiable problem since the relationship

$$E[Y_1 - Y_0] = E[Y_1] - E[Y_0] = E[Y|X = 1] - E[Y|X = 0]$$

does not hold in general, unless the treatment-outcome relationship is unconfounded. The causal contrast Δ is estimable from the data, however, if the assumption of strong ignorability (also known as the no unmeasured confounder or exchangeability assumptions) is invoked. This assumption states that conditionally on C , the treatment is independent of the counterfactual outcomes, i.e. $(Y_1, Y_0) \perp X|C$. Then

$$\begin{aligned} E[Y_1 - Y_0] &= E_C[E[Y_1|C]] - E_C[E[Y_0|C]] \\ &= E_C[E[Y_1|X = 1, C]] - E_C[E[Y_0|X = 0, C]] \\ &= E_C[E[Y|X = 1, C]] - E_C[E[Y|X = 0, C]]. \end{aligned} \quad (1)$$

In addition to the strong ignorability and consistency assumptions, the problem is well formulated only if one assumes that both treatments can be observed for all levels of confounders C , that is $0 < P(X = 1|C) < 1$ (positivity assumption).

A common approach to estimate the causal effect of the treatment, referred to herein as the regression approach, attempts to directly estimate Δ from formulation (1). That is, the analyst first specifies a model for $E[Y|X, C]$ parameterized by β , namely the outcome model. Then the estimation of Δ requires the estimation of β using a least-squares fit, and the averaging of the estimated conditional effect of treatment X over the observed distribution of C (the regression estimator $\hat{\Delta}_R$ is formally defined in Section 3.2.1). Most of the time a linear model that only includes the treatment as a main effect is specified for $E[Y|X, C]$. In this case, Δ simply corresponds to β_X , the parameter associated to the treatment variable X in $E[Y|X, C]$, and the averaging step can be avoided: $\Delta = E_C[E[Y|X = 1, C]] - E_C[E[Y|X = 0, C]] = E_C[\beta_X] = \beta_X$. More generally, estimating Δ falls into the problem of estimating *average effects* (Gelman and Pardoe, 2007; Liu and Gustafson, 2008). Perhaps more frequently in the medical literature, estimating Δ is referred to as estimating the marginal - rather than conditional - effect of the treatment.

Define the propensity score as the probability of treatment given observed covariates, $e(C) = Pr(X = 1|C)$. Assume for the moment that $e(\cdot)$ is known to the investigator. Then, without further modelling assumptions, one can simply construct an unbiased estimator of Δ using inverse probability weighting (IPW):

$$\hat{\Delta}_{IPW} = n^{-1} \sum_{i=1}^n \frac{Y_i X_i}{e(C_i)} - n^{-1} \sum_{i=1}^n \frac{Y_i (1 - X_i)}{1 - e(C_i)}. \quad (2)$$

In practice, the propensity is usually modelled using a logistic model, and $e(C)$ is replaced by $e(C, \hat{\alpha})$ in (2), where α parameterizes the propensity model. In fact, the ability to model the propensity correctly is important as $\hat{\Delta}_{IPW}$ is inconsistent otherwise.

There is a rich framework surrounding IPW for estimating causal effects. Robins and co-authors have shown that (2) is one particular instance among a class of semiparametric estimators that can be used to estimate Δ . Members of this class are known to have differing performance: the stabilized version

$$\hat{\Delta}_S = \left(\sum_{i=1}^n \frac{X_i}{e(C_i, \hat{\alpha})} \right)^{-1} \sum_{i=1}^n \frac{X_i Y_i}{e(C_i, \hat{\alpha})} - \left(\sum_{i=1}^n \frac{1 - X_i}{1 - e(C_i, \hat{\alpha})} \right)^{-1} \sum_{i=1}^n \frac{(1 - X_i) Y_i}{1 - e(C_i, \hat{\alpha})},$$

is usually preferred to (2) and its use is common in practice. However, more sophisticated estimators are also possible. In particular, the one identified as locally efficient among this class is

$$\begin{aligned} \hat{\Delta}_{DR} = & n^{-1} \sum_{i=1}^n \frac{X_i Y_i - (X_i - e(C_i, \hat{\alpha})) m_1(C_i, \hat{\beta}_1)}{e(C_i, \hat{\alpha})} \\ & - n^{-1} \sum_{i=1}^n \frac{(1 - X_i) Y_i + (X_i - e(C_i, \hat{\alpha})) m_0(C_i, \hat{\beta}_0)}{1 - e(C_i, \hat{\alpha})}, \end{aligned} \quad (3)$$

where $m_x(C, \beta_x) = E[Y_x|C]$, $x = 0, 1$ (Robins *et al.*, 1994; Lunceford and Davidian, 2004). This estimator is said to have the double-robustness (DR) property in that $\hat{\Delta}_{DR}$ remains consistent if either i) the propensity e is correctly specified but the two outcome models m_0 and m_1 are not or ii) m_0 and m_1 are correctly specified but e is not, although it need no longer be most efficient in these cases. The DR property is *a priori* appealing since it offers increased protection against biased estimation of the marginal effect Δ . While improved doubly-robust estimators have been considered more recently (e.g., Tan, 2007, 2008), we focus in this work exclusively on estimator (3) which, in our opinion, constitutes a point of reference in the causal inference framework.

2.2 Effect-Measure Modification

Equation (1) suggests that a principled analyst might approach the problem of estimating Δ by modelling the marginals $Y_0|C$ and $Y_1|C$. If one is willing to make symmetric assumptions about $E[Y_0|C]$ and $E[Y_1|C]$ (which seems reasonable) then this corresponds to a form such as

$$E(Y|X, C) = (1 - X)m(C, \beta_0) + Xm(C, \beta_1). \quad (4)$$

Model (4) allows the confounders to affect the counterfactual outcomes differently while assuming a common form m for the expectation being modelled. This formulation of the causal problem thus suggests that interaction terms involving X and C should be considered when specifying $E(Y|X, C)$. When such effect-modification exists, it might be preferable to present stratum-specific estimates, although marginal effects, such as (1), are often valuable summaries. From an interpretation perspective, it is also important to be clear about the population over which the average is applied since, as treated and untreated differ by the distribution of confounders, the marginal treatment effect consequently also differs among these two subpopulations (Stürmer *et al.*, 2006b; Lunt *et al.*, 2009).

Models allowing confounders to also be effect-measure modifiers are not commonly seen in practice. However, it is sensible to believe that a non-negligible bias in the estimate can result if the treatment-confounder interactions are wrongly omitted from the model. Of course, the magnitude and direction of the bias are likely to depend on many factors, such as the strength and direction of the interactions, the propensity for treatment within the levels of the confounders, the prevalence of the confounders in the population. Liu and Gustafson (2008) examine the estimation of marginal effects in linear regression models and identify situations where omitting interaction terms may not have a bad effect on the estimation. These authors show that robustness to this type of misspecification arises when the covariates are either mutually independent or multivariate normal, situations which, by default, do not apply here (the treatment indicator is binary and confounders are by definition not independent of treatment). Greenland (1982) concludes, in a specific context for treatment-effect estimation, that the bias observed by assuming erroneously the homogeneity of effect is prone to be small unless large effect-modification and confounding is present.

In a more general context, while higher-order terms such as interaction terms (perhaps more appropriately product terms) and nonlinear trend terms should be included in statistical models for the sake of realistic effect estimation, the power to detect such terms may be too small in practice (Hernán *et*

al., 2008). Hence these higher-order terms are often left out statistical models, and consequently estimators of main effects may be inconsistent (Greenland, 2009).

3 Methods

3.1 Objective & Scenarios

We compared the performance of the regression estimator for Δ based on representation (1), $\hat{\Delta}_R$, versus the doubly-robust estimator $\hat{\Delta}_{DR}$. More specifically, our goal was to examine to what extent outcome model misspecification affects the mean square error (MSE) of these estimators. We also considered the performance of the IPW estimator $\hat{\Delta}_S$ to provide another benchmark for assessing the impact of misspecifying the outcome model when using $\hat{\Delta}_{DR}$.

In our scenarios, we assumed that all confounders were measured and included in both the propensity and outcome models. Because the impact of misspecifying the outcome model was of primary interest, we assumed a correctly specified propensity model. The logit of the conditional probability of being treated given C was

$$\text{logit}(e(C, \alpha)) = \text{logit}(\text{Pr}(X = 1|C)) = \alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3, \quad (5)$$

where C_1 , C_2 and C_3 are defined below. For all but one scenario we assumed $C_1 \sim \text{Ber}(0.5)$, $C_2^*, C_3^* \sim N(0, 1)$ and C_1, C_2^*, C_3^* independent. Otherwise (C_2^*, C_3^*) followed a bivariate standard normal with correlation coefficient $\rho = 0.3$. Because the coefficients in regression models are sensitive to the scale of the inputs, we standardized the continuous covariates C_2^*, C_3^* and instead considered $C_2 = C_2^*/2$ and $C_3 = C_3^*/2$. This standardization ensures that the continuous covariates are on the same scale as the dichotomous covariate C_1 (Gelman, 2008). Finally, since the performance of the estimators across the parameter space was of interest, we regarded the parameters of the propensity model α_i ($i = 0, \dots, 3$) as independent realizations from a uniform prior distribution π_α on the interval $(-\log 4, \log 4)$.

The outcome model misspecification was introduced by omitting some interactions or non-linear effects. The conditional distribution of Y was normal with mean $E[Y|X, C]$ and residual variance $\sigma^2 = 1$. In Table 1 we present the seven scenarios of model misspecification for $E[Y|X, C]$ examined, representing instances where either an interaction between two confounders, a quadratic effect of a confounder or an interaction between the treatment and a confounder is possibly missing. Moreover, we also investigate the effect of

such omissions for different types of confounders (dichotomous versus continuous, independent versus dependent). The prior distribution of the parameters of the outcome models, π_β , was the product of independent uniforms on the interval $(-3, 3)$. Note that we used the same generated α parameter values for all scenarios and the same β parameter values for true outcome models having the same dimension.

Table 1: Covariate inclusion for true (T) and fitted (F) outcome regression models (by scenario)

Scenario	Baseline	C_1C_2	C_2C_3	C_2^2	XC_1	XC_2	XC_2^2
T ₁	×	×					
F ₁	×						
T ₂	×		×				
F ₂	×						
T _{3*}	×		×				
F _{3*}	×						
T ₄	×			×			
F ₄	×						
T ₅	×				×		
F ₅	×						
T ₆	×					×	
F ₆	×						
T ₇	×			×		×	×
F ₇	×			×		×	

NOTE: All models include an intercept term, the treatment indicator X and the three confounders C_1 , C_2 and C_3 as main effects (baseline covariates). *: In Scenario 3, the covariates C_2 and C_3 are dependent (as opposed to Scenario 2).

To our knowledge, performance of regression and IPW estimators has not been previously investigated in the “random parameter value” manner employed here. We take this approach for a number of reasons. First, we anticipate that our qualitative findings are less sensitive to the choice of prior distributions than findings based on selecting fixed points in the parameter space. Second, this formulation is conducive to examining how estimator performance varies smoothly with a particular feature of the data-generating mechanism, in a sense which is aggregated over other features. For instance we can readily examine how estimator performance varies as a function of the propensity

scores, or with the detectability of regression model misspecification. Third, along the usual lines in experimental design, interpolating performance at a given point in the parameter space from performances at nearby sampled points is easier than extrapolating results from a small, fixed set of parameters to the remainder of the parameter space.

The type of model misspecification we considered in this study is of similar flavour to that of Kang and Schafer (2007), although these authors did not consider the case arising from missing treatment-confounder interactions. However, it is different from Lunceford and Davidian (2004) who focused on the omission of a confounder that is nevertheless included in the propensity model. As these authors suggest, this type of model misspecification is rather extreme and the bias occurring when a confounder is omitted is well known. Lunceford and Davidian’s scenario in fact mimics the situation where a covariate is mistakenly considered as a pure predictor of exposure (predictor of exposure but not predictor of outcome) and thus is not included in the outcome model. However, many empirical investigations have suggested the practice of including only confounders and pure predictors of outcome in the propensity model for IPW and PS (e.g., Austin *et al.*, 2007; Brookhart *et al.*, 2006).

We performed some inspections to assess the degree of misspecification of the models fitted for Scenarios 1, 4, 5 and 7. We generated 2000 data sets of size 100, each based on parameter values α_k and β_k ($k = 1, \dots, 2000$) drawn from the prior distributions π_α and π_β . For each scenario and data set, we tested for the null effect of the omitted term and for the null effect of all six terms considered in Table 1 altogether (excluding those already included in the fitted model, if applicable). When testing for the null effect of the omitted term only, we obtained for Scenarios 1, 4, 5 and 7, 74%, 81%, 71% and 47% of the p-values smaller than 0.05, respectively. When testing for the null effect of all terms together, we obtained 62%, 71%, 59% and 34% of the p-values smaller than 0.05. The mean adjusted R^2 for the fitted model in Scenarios 1, 4, 5 and 7 was 0.69 (sd= 0.14), 0.66 (sd= 0.14), 0.71 (sd= 0.14) and 0.75 (sd= 0.11), respectively. Both p-values and adjusted R^2 were globally larger for the two models truly missing some interaction terms between treatment and confounders.

3.2 Large-Sample Expressions

It is well known that, from a decision-theoretic perspective, the MSE is a meaningful quantity to compare different decision rules for point estimation, since it is the frequentist risk under the squared error loss (Bernardo and Smith, 1994). Unlike most existing studies, we assessed the performance of

the estimators based on the MSE derived from large-sample theory instead of repeated data-set simulation. We are aware that asymptotics may be “slow to kick in” for IPW estimators under some circumstances and that ad hoc weight truncation might improve their MSE in practice (e.g. Cole and Hernán, 2008); hence we are viewing our results as being useful for making contrasts between different estimators under different circumstances, rather than necessarily giving excellent numerical approximation to finite-sample performance in all settings. Given that computational resources are limited, we believe that more information on the estimators can be gleaned by exploring approximate (due to asymptotics) performance comprehensively across the parameter space and across sample size rather than determining approximate (due to simulation error) performance at a few selected parameter values and a few selected sample sizes.

In this section we present the large sample expressions for the bias and variance of $\hat{\Delta}_R$, $\hat{\Delta}_S$ and $\hat{\Delta}_{DR}$ required for the computation of the MSE.

3.2.1 Regression Estimator $\hat{\Delta}_R$

Let $W = (X, C_1, \dots, C_p)^T$ comprise the set of the treatment variable and confounders used for representation (4) and let $T(W) = \{T_1(W), \dots, T_q(W)\}^T$ be the true predictors of Y , so that

$$Y = T^T \beta + \varepsilon,$$

where ε is independent of W , with $E[\varepsilon] = 0$ and $\text{Var}[\varepsilon] = \sigma^2$. Similarly, let $S(W) = \{S_1(W), \dots, S_r(W)\}^T$ be the fitted predictors. Further, let us define $\tilde{T}$ and $\tilde{S}$ as

$$\tilde{T} = \{T(1, C_1, \dots, C_p) - T(0, C_1, \dots, C_p)\}^T$$

and

$$\tilde{S} = \{S(1, C_1, \dots, C_p) - S(0, C_1, \dots, C_p)\}^T.$$

Following Liu and Gustafson (2008), we use double-struck font to indicate design matrices and outcome vector and define

$$\hat{\Delta}_R = n^{-1} \mathbb{I}_n^T \tilde{\mathbb{S}} (\mathbb{S}^T \mathbb{S})^{-1} \mathbb{S}^T \mathbb{Y}.$$

When the true outcome model is fitted (i.e. $S \equiv T$) we obtain

$$n^{1/2} \left\{ \hat{\Delta}_R - \Delta(\beta) \right\} \xrightarrow{D} N(0, v_R(\alpha, \beta)),$$

where $\Delta(\beta)$ is the true marginal treatment effect

$$\Delta(\beta) = E \left[\tilde{T}^T \right] \beta$$

and

$$v_R(\alpha, \beta) = \sigma^2 E \left[\tilde{T}^T \right] \{ E [TT^T] \}^{-1} E \left[\tilde{T} \right] + \beta^T \text{Var} \left[\tilde{T} \right] \beta.$$

These results are generalized as follows in the case of a misspecified model:

$$n^{1/2} \left\{ \hat{\Delta}_R - \Delta^*(\alpha, \beta) \right\} \xrightarrow{D} N(0, v_R^*(\alpha, \beta)),$$

where

$$\Delta^*(\alpha, \beta) = E \left[\tilde{S}^T \right] \{ E [SS^T] \}^{-1} E [ST^T] \beta$$

and

$$\begin{aligned} v_R^*(\alpha, \beta) = & \sigma^2 E \left[\tilde{S}^T \right] \{ E [SS^T] \}^{-1} E \left[\tilde{S} \right] \\ & + E \left[\tilde{S}^T \right] \{ E [SS^T] \}^{-1} E \left[S(G^T \beta)^2 S^T \right] \{ E [SS^T] \}^{-1} E \left[\tilde{S} \right] \\ & + \beta^T E [TS^T] \{ E [SS^T] \}^{-1} \text{Var} \left[\tilde{S} \right] \{ E [SS^T] \}^{-1} E [ST^T] \beta \\ & + 2\beta^T E [TS^T] \{ E [SS^T] \}^{-1} \text{Cov} \left[\tilde{S}, SG^T \beta \right] \{ E [SS^T] \}^{-1} E \left[\tilde{S} \right], \end{aligned}$$

where $G = T - E [TS^T] \{ E [SS^T] \}^{-1} S$.

The bias in the estimate arising from specifying S differently than T in the regression outcome model is thus $b_R = \Delta^* - \Delta$.

3.2.2 Inverse Probability Weighted Estimators $\hat{\Delta}_S$ and $\hat{\Delta}_{DR}$

Since we used the true propensity model in our experimental design, the large-sample bias of $\hat{\Delta}_S$ is zero. This is also true for $\hat{\Delta}_{DR}$, even though the fitted outcome model is misspecified in each scenario of Table 1. Therefore under suitable regularity conditions

$$n^{1/2} \left\{ \hat{\Delta}_{S,DR} - \Delta \right\} \xrightarrow{D} N(0, v_{S,DR}(\alpha, \beta)).$$

The expression for the variance of $\hat{\Delta}_S$ with the propensity modelled via a logistic regression is taken directly from Lunceford and Davidian (2004) and is given by

$$v_S(\alpha, \beta) = E \left(\frac{(Y_1 - \mu_1)^2}{e(C, \alpha)} \right) + E \left(\frac{(Y_0 - \mu_0)^2}{1 - e(C, \alpha)} \right) - H_\alpha^T E_{\alpha\alpha}^{-1} H_\alpha, \quad (6)$$

where the vector H_α is defined as

$$H_\alpha = E \left\{ \left(\frac{Y_1 - \mu_1}{e(C, \alpha)} + \frac{Y_0 - \mu_0}{1 - e(C, \alpha)} \right) e_\alpha(C) \right\}$$

with $e_\alpha(C) = \partial/\partial\alpha \{e(C, \alpha)\}$, and the matrix $E_{\alpha\alpha}$ as

$$E_{\alpha\alpha} = E \left(\frac{e_\alpha(C) e_\alpha(C)^T}{e(C, \alpha)(1 - e(C, \alpha))} \right).$$

Note that (6) takes into account that the propensity is estimated; the variance of $\hat{\Delta}_S$ when the propensity is known is obtained by dropping the last term.

Lunceford and Davidian (2004) also present the expression for the variance of $\hat{\Delta}_{DR}$, but in the particular case of a correctly specified outcome model. In the present context, the variance of $\hat{\Delta}_{DR}$ when the outcome model is misspecified is of primary interest. Indeed, one realistic assumption to make is that the analyst is equally proficient at specifying the outcome model when using $\hat{\Delta}_R$ or $\hat{\Delta}_{DR}$. We first derive the variance of $\hat{\Delta}_{DR}$ in a specialized case. Assume that in (3), the propensity model is correct and the parameters are known. Moreover also assume that the parameters in the fitted outcome models $m^*(C, \beta_x)$ $x = 0, 1$ are known, in the sense that the large sample limit of the estimator (as obtained by $\{E[SS^T]\}^{-1} E[ST^T] \beta$) is known. Then some algebra shows that the variance of $\hat{\Delta}_{DR}$ can be written as

$$v_{DR}^*(\alpha, \beta) = Var[m(C, \beta_1) - m(C, \beta_0)] + E \left[\frac{1}{e(C)(1 - e(C))} \right] \sigma^2 + term, \quad (7)$$

where *term* is the expression for the extra variance arising from specifying incorrectly the outcome model (note that *term* is almost surely non negative and equal to zero when $m^*(C, \beta_x) \equiv m(C, \beta_x)$):

$$term = E \left\{ \left(\sqrt{\frac{1 - e(C)}{e(C)}} [m(C, \beta_1) - m^*(C, \beta_1)] + \sqrt{\frac{e(C)}{1 - e(C)}} [m(C, \beta_0) - m^*(C, \beta_0)] \right)^2 \right\}. \quad (8)$$

The decompositions (7-8) of the variance of the DR IPW estimator indicates that there is no free lunch: its performance may be hampered by a wrong specification of the outcome model, particularly when there is a strong indication to treatment - or to no treatment, e.g. when $e(C)$ is very large or

small. This behaviour will be empirically observed in our results. More realistically, the parameters of both propensity and outcome models are estimated. When $m^* \equiv m$, the variance of the DR estimator is the same whether the parameters of the models are estimated or not. This is not generally the case when the true and fitted models differ. Therefore we also derive $v_{DR}^*(\alpha, \beta)$ in this case using M-estimation theory (a good introduction is given in Stefanski and Boos, 2002). Define $\theta^T = (\mu_1, \mu_0, \alpha^T, \beta^T)$, then $\hat{\theta}_{DR}$ is the solution to $\sum_{i=1}^n \psi_i(\hat{\theta}_{DR}) = 0$, where ψ_i is the following set of stacked estimating equations

$$\psi_i = \begin{pmatrix} \psi_{1,i} \\ \psi_{2,i} \\ \psi_{3,i} \\ \psi_{4,i} \end{pmatrix} = \begin{pmatrix} \frac{X_i Y_i - (X_i - e(C_i, \alpha))m^*(C_i, \beta_1)}{e(C_i, \alpha)} - \mu_1 \\ \frac{(1 - X_i)Y_i + (X_i - e(C_i, \alpha))m^*(C_i, \beta_0)}{1 - e(C_i, \alpha)} - \mu_0 \\ \frac{X_i - e(C_i, \alpha)}{e(C_i, \alpha)(1 - e(C_i, \alpha))} e_\alpha(C_i) \\ \frac{(Y_i - \beta^T S_i)S_i}{e(C_i, \alpha)(1 - e(C_i, \alpha))} \end{pmatrix},$$

S is the vector of the fitted predictors of outcome as defined in Section 3.2.1 and $m^*(C, \beta_x)$ is the fitted model for $E[Y_x|C]$, $x = 0, 1$. Writing $\partial/\partial\alpha \{e(C, \alpha)\} = e(C, \alpha)(1 - e(C, \alpha))P$, where P is the vector of predictors of the propensity, $\psi_{3,i}$ simplifies to $(X_i - e(C_i, \alpha))P_i$. Then the variance of $\hat{\Delta}_{DR}$ is given by

$$v_{DR}^*(\alpha, \beta) = V_{11}(\theta) + V_{22}(\theta) - 2V_{12}(\theta)$$

where $V(\theta) = A(\theta)^{-1}B(\theta)\{A(\theta)^{-1}\}^T$, $B(\theta) = E[\psi(\theta)\psi(\theta)^T]$,

$$\begin{aligned} A(\theta) &= E \left[-\frac{\partial}{\partial \theta^T} \psi(\theta) \right] \\ &= E \begin{pmatrix} 1 & 0 & \frac{(Y - m^*(C, \beta_1))X}{e(C, \alpha)^2} e_\alpha(C) & \frac{(X - e(C, \alpha))}{e(C, \alpha)} m_{\beta_1}^*(C) \\ 0 & 1 & \frac{-(Y - m^*(C, \beta_0))(1 - X)}{(1 - e(C, \alpha))^2} e_\alpha(C) & \frac{-(X - e(C, \alpha))}{(1 - e(C, \alpha))} m_{\beta_0}^*(C) \\ 0 & 0 & e(C, \alpha)(1 - e(C, \alpha))PP^T & 0 \\ 0 & 0 & 0 & SS^T \end{pmatrix}, \end{aligned}$$

and $m_{\beta_x}^*(c) = \partial/\partial\beta_x \{m^*(C, \beta_x)\}$.

4 Results

For each scenario we generated 2000 pairs of vector of parameters (α_k, β_k) ($k = 1, \dots, 2000$) for the true propensity and outcome models from the priors

π_α and π_β . For every pair, we then computed the quantities Δ^* , v_R^* , v_{DR}^* and v_S resulting from the definition of the true and fitted outcome models in a given scenario. Computing Δ^* , v_R^* , v_{DR}^* and v_S required numerically approximating all expectation terms involved in their expressions. We calculated the conditional expectations $E_{X|C}[\cdot]$ exactly wherever required, and relied on Monte Carlo integration for estimating the expectations with respect to C , using 10,000 independent samples for (C_1, C_2, C_3) . Note that the same sequence of random numbers was used to generate the confounders C_1, C_2, C_3 in all scenarios. We used the R (version 2.6.1) programming language to perform the computations.

4.1 Examining the Variance of $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$ in Presence of Outcome Model Misspecification

While our ultimate interest is in MSE, we first looked at the variance of $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$ in presence of outcome model misspecification. In Figure 1 we present scatterplots comparing the large sample value of the variance between the regression and doubly-robust estimators, $n^{-1}v_R^*$ versus $n^{-1}v_{DR}^*$ (more scatterplots are presented in Figure 7 of the web appendix). Since both variances decrease with a common n^{-1} factor with increasing sample size, we present results only for the case $n = 1$. For all scenarios of misspecification but Scenario 5 (interaction term XC_1 missing), $\hat{\Delta}_R$ exhibited smaller variance than $\hat{\Delta}_{DR}$ uniformly across parameter values (α_k, β_k) , $k = 1, \dots, 2000$. We also compared the large sample value of the variance between the stabilized and doubly-robust estimators (v_S versus v_{DR}^*) (see Figure 8 in the web appendix). Recall that the doubly-robust estimator is guaranteed to have smaller variance than the stabilized estimator only when both the propensity and outcome models are correctly specified, and thus an incorrect specification of the outcome model may no longer make $\hat{\Delta}_{DR}$ efficient. In the scenarios examined, the misspecification of the outcome model did not have a dramatic impact on the performance of the doubly-robust estimator. Indeed, $\hat{\Delta}_{DR}$ showed smaller variance than $\hat{\Delta}_S$ for most parameter values. This situation occurred in the following proportions for Scenarios 1, 4, 5 and 7, respectively: 0.87, 0.82, 0.95, 0.93. Interestingly, the performance of $\hat{\Delta}_{DR}$ (as opposed to $\hat{\Delta}_S$) was best when the misspecification occurred through an omitted treatment-confounder interaction term (Scenarios 5 and 7), cases where the p-values for detectability of the omitted term and adjusted R^2 were the largest. To provide a point of reference for the results we compared v_R and v_{DR} , and v_{DR} and v_S for a correctly specified model with an XC_1 effect (the true model in Scenario 5). It is well known from theory that $\hat{\Delta}_R$ performs better than $\hat{\Delta}_{DR}$ when the true

outcome model is fitted to data. As expected, v_R was smaller than v_{DR} for every parameter value, as was v_{DR} compared to v_S (see Figure 9 in the web appendix).

Figure 1: Variance of the regression and doubly robust estimators for Scenarios 1 and 5 (v_R^* vs v_{DR}^* on log scale). Display: top to bottom, respectively.

4.2 Aggregate Performances for $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$

More results are given for Scenarios 4 and 5 in Figures 2 and 3; results for all scenarios are given in the web appendix (see Figures 11 to 14). Specifically, in Figures 2-3 (a) and (b) we present intensity plots for the large-sample mean-square errors (MSE) of $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$. In Figures 2-3 (c) we present intensity plots for the bias and variance of $\hat{\Delta}_R$. Finally, the proportion of MSE due to the bias of $\hat{\Delta}_R$ is illustrated in Figures 2-3 (d). All these quantities are pictured as a joint function of the expected value of the product of the propensity score for treated and untreated, $E[e(C)(1 - e(C))]$, and of the absolute value of the outcome model parameter of the missing covariate, $|\beta_m|$. The MSE of $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$ as a function of sample size is given by $MSE_R^*(n) = (\Delta^*(\alpha_k, \beta_k) - \Delta(\beta_k))^2 + n^{-1}v_R^*(\alpha_k, \beta_k)$ and $MSE_{DR}^*(n) = n^{-1}v_{DR}^*(\alpha_k, \beta_k)$, for $k = 1, \dots, 2000$. To produce the plots, we considered five equally-sized categories based on the cutoffs for the 20, 40, 60 and 80th per-

centiles of $E[e(C)(1 - e(C))]$ and five equally-sized categories based on the cutoffs for the 20, 40, 60 and 80th percentiles of $|\beta_m|$. These levels corresponded to $Q_1 : [0.110 - 0.156)$, $Q_2 : [0.156 - 0.171)$, $Q_3 : [0.171 - 0.184)$, $Q_4 : [0.184 - 0.200)$, $Q_5 : [0.200 - 0.246)$ for $E[e(C)(1 - e(C))]$ and $Q_1 : (0, 0.627)$, $Q_2 : [0.627 - 1.240)$, $Q_3 : [1.240 - 1.818)$, $Q_4 : [1.818 - 2.385)$, $Q_5 : [2.385 - 3)$ for $|\beta_m|$. We then obtained *average* performance of the estimators within each combination of the $(E[e(C)(1 - e(C))], |\beta_m|)$ quintiles.

The rationale behind this analysis was to investigate the behaviour of the estimators in presence of different degrees of outcome model misspecification and indication to treatment. We choose $E[e(C)(1 - e(C))]$ as an aggregation criterion for the following reasons. First we empirically observed that the performance of $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$ varied smoothly over the levels of $E[e(C)(1 - e(C))]$. In this regard, the expression for the variance of $\hat{\Delta}_{DR}$ presented in (7) is particularly enlightening since it directly shows that the variance largely depends on the extent of distributional balance between the treatment groups. We chose the quantity $E[e(C)(1 - e(C))]$ since it represents a nice measure of the balance of the treated versus non-treated individuals in the population. The product $e(c)(1 - e(c))$ lies between 0 and 0.25 and is maximum and when the propensity to treatment is equal to 0.5 (given covariates values c) and thus $E[e(C)(1 - e(C))]$ achieves its maximum value when the propensity to treatment and to absence of treatment are equal and are independent of C . Both propensities play equally important roles in the definition and properties of the estimators examined. Therefore we also aimed for a “symmetrical” criterion, which is the case here by commutativity of the multiplication. This, in turn, enabled us to consider a unidirectional scaling for our criterion, where larger values of $E[e(C)(1 - e(C))]$ reflect increased balance between the distributions of treated versus untreated individuals.

For purposes of interpretation of the intensity plots, note that the quantities of interest (MSE, bias, variance) are normalized to the grey level scale (0: white to 1: black) presented in Figure 4 (Figure 10 in web appendix). More precisely, the normalization is achieved by dividing the average measures by the maximum value observed over quintile combinations of a given intensity plot.

Unsurprisingly, we first observed across all scenarios that the variance of $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$ increased with the magnitude of $|\beta_m|$ but decreased along with $E[e(C)(1 - e(C))]$. The same behavior was observed for the bias of $\hat{\Delta}_R$. The estimators thus exhibited worst MSE when simultaneously in the most problematic quintiles for $|\beta_m|$ and $E[e(C)(1 - e(C))]$, that is where the misspecification of the outcome model and the imbalance between treated and untreated is at their maximum.

Figure 2: Scenario 4 (C_2^2 misspecification): Intensity plots as a function of 0-20, 20-40, 40-60, 60-80, and 80-100th percentiles of $E[e(C)(1 - e(C))]$ (x-axis) and of $|\beta_m|$ (y-axis) (based on 2000 different (α, β)). (a) Top to bottom: Average $MSE_R^*(n)$, $n = 1, 100, 500, 1000, 2000$ (b) Top to bottom: Average $MSE_{DR}^*(n)$, $n = 1, 100, 500, 1000, 2000$ (c) Top to bottom: Average b_R^2 , v_R^* . (d) Top to bottom: Average $b_R^2 / MSE_R^*(n)$, $n = 1, 100, 500, 1000, 2000$. **Notes:** for better comparisons between $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$, the grey levels of plots appearing in subfigures (a) and (b) are normalized using a common scale (for each sample size separately). Plots in (c) are drawn on their own individual scale. Plots in (d) are left unnormalized since $b_R^2 / MSE_R^*(n) \in [0, 1]$. See Figure 4 for grey level scale.

Figure 3: Scenario 5 (XC_1 misspecification). Layout and legend as per Figure 2.

Figure 4: Grey level scale from 0.05 to 0.95 by 0.10 (0: white - 1: black).

In order to better understand the variance results presented earlier, we also obtained intensity plots for the difference in variances between $\hat{\Delta}_{DR}$ and $\hat{\Delta}_R$ (see Figure 15 in the web appendix). These plots provide an indication of the magnitude of the difference $v_{DR}^* - v_R^*$ in the various the regions of the $(|\beta_m|, E[e(C)(1 - e(C))])$ space. For Scenarios 1, 4 and 7, the differences behaved similarly, where the largest ones were observed as we move towards the highest and lowest quintiles of $|\beta_m|$ and $E[e(C)(1 - e(C))]$, respectively. We observed in Figure 1 that Scenario 5 behaved differently than the other scenarios. This is the case here too, where the doubly robust estimator possessed larger variance than the regression estimator towards the lowest quintiles of both $|\beta_m|$ and $E[e(C)(1 - e(C))]$, that is, when the imbalance in treatment is large but the misspecification small.

Overall, the performance of $\hat{\Delta}_R$ (from MSE standpoint) as opposed to $\hat{\Delta}_{DR}$ deteriorated for the larger sample sizes $n = 1000$ and 2000 , since the bias b_R did not vanish as the sample size increased. The performance of $\hat{\Delta}_R$ was comparatively poorer for Scenario 5, where interaction XC_1 is wrongly omitted from the outcome model. In this particular case, the MSE of $\hat{\Delta}_{DR}$ was smaller than that of $\hat{\Delta}_R$ for most quintile combinations as early as $n = 100$. The most favourable scenario for $\hat{\Delta}_R$ was Scenario 4 (quadratic term C_2^2 missing) where it had worse performance than $\hat{\Delta}_{DR}$ only for a few quintile combinations at the largest sample size.

The bias in the estimates from regression with a misspecified outcome model was significantly larger when treatment-confounder interactions were omitted. This was particularly apparent in Scenario 5 where the performance of $\hat{\Delta}_R$ appeared to be primarily affected by its large amount of bias, in comparison to its variance. We conjecture that one reason explaining why the results were worst for the scenarios that omit interaction terms between the treatment and the confounders is that the outcome model may be more likely to yield dissimilar patterns of bias for $Y_0|C$ and $Y_1|C$ in these cases.

We observed similar results when considering different types of confounders (continuous versus dichotomous and independent versus dependent), as in Scenarios 2, 3 and 6 (results not shown). Both $\hat{\Delta}_{DR}$ and $\hat{\Delta}_R$ were slightly negatively affected by non independent confounders (Scenario 3). However these estimators were differently affected by continous confounders, where a slight increase and decrease in MSE was observed for $\hat{\Delta}_{DR}$ and $\hat{\Delta}_R$, respectively (Scenarios 2 and 6).

4.3 Δ_R 's Coverage Performance

We now report some indication of the coverage performance of $\hat{\Delta}_R$ when the outcome model is misspecified. A useful rule-of-thumb is that the performance

of interval estimates and test statistics begins to deteriorate when the (absolute value of the) bias exceeds about 40 percent of its standard deviation (Kang and Schafer, 2007). For each scenario, we calculated the number of times (out of 2000) where this situation occurred, as a function of sample size. The results are presented in Table 2. These results indicate that the regression estimator coverage performance was impaired overall, most particularly for Scenarios 5-7 which all wrongly omit some treatment-confounder interaction terms.

Table 2: Coverage results for the regression estimator. Number of occurrences where $|\Delta^* - \Delta| > 0.4\sqrt{v_R^*/n}$, by scenario and sample size (out of 2000)

Scenarios	$n = 1$	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$
1	0	2	100	215	357
2	0	0	72	179	308
3	0	15	132	239	356
4	0	0	148	264	407
5	0	533	1012	1165	1296
6	0	517	1004	1171	1320
7	0	25	377	574	770

4.4 Assessing the Direct Impact of Model Misspecification on $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$

To better understand the effect of misspecification on the regression and doubly-robust estimators, we then obtained, for Scenarios 1, 4, 5 and 7, performance measures when the true model was fitted to data. A complementary objective of the study was indeed to quantify the additional MSE due to misspecification for each estimator, that is, to assess the difference in MSE when the true and misspecified models were fitted to data, for both $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$. For Scenario 1 for example, we were interested in assessing the difference in MSE when 1) the term C_1C_2 was wrongly omitted in the fitted model and 2) when the term C_1C_2 was correctly included in the fitted model. Using this approach thus permitted us to distinguish between the contribution of the “baseline” MSE of the estimators and the error directly induced by the misspecification of the outcome model. Results are presented for Scenarios 1 and 5 in Figures 5 and 6 (see Figures 16 to 19 in the web appendix for all scenarios).

Across all scenarios, the superiority of $\hat{\Delta}_{DR}$ in terms of MSE arrived at smaller sample sizes when we only considered the supplemental MSE due to

misspecification. This is not entirely surprising since the regression estimator possesses smaller variance than the doubly-robust estimator when the true model is fitted (that is at baseline). For instance, for Scenario 1, the additional MSE of $\hat{\Delta}_{DR}$ compared to $\hat{\Delta}_R$ at $n = 2000$ was smaller for all but one combination of quintiles of $E[e(C)(1 - e(C))]$ and of $|\beta_m|$, whereas the MSE of $\hat{\Delta}_{DR}$ was still greater than the one for $\hat{\Delta}_R$ for many quintile combinations when the baseline MSE was accounted for. Less marked differences were however observed for Scenario 5, where the performance of $\hat{\Delta}_R$ was already much impaired at the smaller sample sizes.

We also examined the impact of misspecification on the additional variance (rather than MSE) of $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$. First, larger (or smaller) differences $v_R^* - v_R$ and $v_{DR}^* - v_{DR}$ occurred in the same regions of the $E[e(C)(1 - e(C))]$ and $|\beta_m|$ quintiles. Across all scenarios, we observed that $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$'s increase in variance depended on the interaction between $E[e(C)(1 - e(C))]$ and $|\beta_m|$, that is a large increase was observed for large $|\beta_m|$ and small $E[e(C)(1 - e(C))]$ simultaneously. This was also the pattern exhibited for the bias of the regression estimator. However, the magnitude of the missing parameter coefficient appeared to be the primary factor affecting the supplemental variance of both estimators in Scenarios 1 and 4. It is interesting to note that in Scenarios 5 and 7 the estimator $\hat{\Delta}_R$ was greatly affected by the value of $|\beta_m|$ especially in the first quintile of $E[e(C)(1 - e(C))]$: although omitting the interaction term appeared to be very detrimental for its variance when $|\beta_m|$ was large, minor gains were seen when $|\beta_m|$ was small. Indeed it appears from Scenarios 5 and 7 that one could induce a minor reduction in the variance of the regression estimator by purposely omitting treatment-confounder interaction terms when the effect-modification is small. Last but not least, while we observed that the estimators were affected by the misspecification in the same regions of the $E[e(C)(1 - e(C))]$ and $|\beta_m|$ quintiles, note that the supplemental variance of $\hat{\Delta}_{DR}$ due to misspecification was uniformly larger than the one for $\hat{\Delta}_R$ in Scenarios 1, 4, 7. This seems to counterbalance (maybe unsatisfactorily) the non-null bias of $\hat{\Delta}_R$ in such situations.

5 Discussion

Our study was designed to investigate the extent to which model misspecification of the outcome model affects the performance of the regression and doubly-robust inverse probability weighted estimators to assess causal treatment effects. While there are some simulation studies in the literature that examine regression and IPW approaches, our work appears to be the first

to employ sampling across the parameter space, and consequently compares methods comprehensively across the space. This seems important, as comparisons based on a few select parameter values can be sensitive to the chosen values. Indeed, intensity plots illustrate considerable variability across parameter values in the performances of the estimators in presence of misspecification. Not so surprisingly, the largest impact of misspecification on the variance and bias of the estimators was found in the most problematic region of the parameter space considered.

Another point of departure between our work and the existing literature involves the emphasis on treatment-confounder interaction terms. One argument for including such terms in regression models is simply conceptual. If one starts with the causal paradigm and thinks about $Y_0|C$ and $Y_1|C$ as arising from two distributions possibly indexed by different sets of parameter values for C , then one is led to including such interaction terms in the regression model for $Y|X, C$. In addition to this, though, our empirical work also suggests that including such terms can be a good modelling strategy. That is, the scenarios where regression performed worst relative to doubly-robust inverse probability weighting involved omitting these terms when they were actually present. Moreover, the detection of treatment-confounder interactions was found to be the most difficult in our study. Note, however, that these results were obtained in a very specific context of treatment effect estimation, that is, a context wherein the effect of a time-independent binary treatment on the mean of a continuous outcome is of interest. The extension of these conclusions to more complex and realistic treatment effect estimation frameworks, needing survival analysis techniques for example, would obviously require further investigation.

While no attempt is made to discriminate between estimators, our results hint that for small-to-moderate sample sizes, the regression estimator is often preferable to the doubly-robust estimator in terms of MSE, even when the outcome model is misspecified. It is also worth mentioning that the MSE calculations were based on an optimistic set-up for the latter estimator, i.e. a correct and very simple propensity model. Importantly, our results suggest that the decrease in performance of the regression estimator with sample size when the model is misspecified almost exclusively comes from its bias, as its variance was almost always smaller than that of the doubly-robust estimator. As a matter of fact, we observed that the regression estimator often exhibited a large amount of bias in proportion to its variance, which negatively affected the coverage performance of the corresponding interval estimator. This seemed to be especially problematic when the omitted term was a treatment-confounder interaction. Although the impact of missed confounders in the

Figure 5: Scenario 1 (C_1C_2 misspecification): Intensity plots as a function of 0-20, 20-40, 40-60, 60-80, and 80-100th percentiles of $E[e(C)(1 - e(C))]$ (x-axis) and of $|\beta_m|$ (y-axis) (based on 2000 different (α, β)). (a) Top to bottom: Average $MSE_R^*(n) - MSE_R(n)$, $n = 1, 100, 500, 1000, 2000$ (b) Top to bottom: Average $MSE_{DR}^*(n) - MSE_{DR}(n)$, $n = 1, 100, 500, 1000, 2000$ (c) Average $v_R^* - v_R$ (d) Average $v_{DR}^* - v_{DR}$. **Notes:** for better comparisons between $\hat{\Delta}_R$ and $\hat{\Delta}_{DR}$, the grey levels of plots appearing within subfigure pairs (a,b) and (c,d) are normalized using a common scale (for each sample size separately when applicable).

Figure 6: Scenario 5 (XC_1 misspecification): Layout and legend as per Figure 5. Bold borders indicate negative value for the quantity of interest.

regression model for treatment effect estimation has been studied extensively, the impact of other forms of model misspecification is less well known. In light of these results, we believe it worthy to further investigate the contexts where the common regression approach is biased.

Web Appendix

Supplemental figures are available in the web appendix
(<http://www.stat.ubc.ca/~gustaf/IJBLGWebAppendix.pdf>).

References

- Austin, P.C., Grootendorst, P., Anderson, G.M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*, **26**(4), 734–753.
- Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, **163**(12), 1149–1156.
- Bernardo, J.M., Smith, A.F.M. (1994). *Bayesian theory*, Chichester: John Wiley & Sons.
- Cole, S.R., Hernán, M.A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, **168**(6), 656–664.
- D’Agostino Jr., R.B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, **17**(19), 2265–2281.
- Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, **27**(15), 2865–2873.
- Gelman, A., Pardoe, I. (2007). Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, **37**(1), 23–51.
- Greenland, S. (1982). Interpretation and estimation of summary ratios under heterogeneity. *Statistics in Medicine*, **1**(3), 217–227.
- Greenland, S. (2009). Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology*, **20**(1), 14–17.

- Hernán, M.A., Alonso, A., Logan, R., Grodstein, F., Michels, K.B., Willett, W.C., Manson, J.E., Robins, J.M. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, **19(6)**, 766–779.
- Hirano, K., Imbens, G. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcome Research Methodology*, **2**, 259–278.
- Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81(396)**, 945–960.
- Kang, J.D.Y., Schafer, J.L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22(4)**, 523–539.
- Liu, J., Gustafson, P. (2008). On average predictive comparisons and interactions. *International Statistical Review*, **76(3)**, 419–432.
- Lunceford, J.K., Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23(19)**, 2937–2960.
- Lunt, M., Solomon, D., Rothman, K., Glynn, R., Hyrich, K., Symmons, D.P.M., Stürmer, T., and the British Society for Rheumatology Biologics Register, the British Society for Rheumatology Biologics Register Control Centre Consortium (2009). Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *American Journal of Epidemiology*, **169(7)**, 909–917.
- Robins, J.M., Rotnitzky, A., Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89(427)**, 846–866.
- Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, **82(398)**, 387–394.
- Rosenbaum, P.R., Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70(1)**, 41–55.
- Stefanski, L.A., Boos, D.D. (2002). The calculus of M-estimation. *The American Statistician*, **56(1)**, 29–38.

- Stürmer, T., Joshi, M., Glynn, R.J., Avorn, J., Rothman, K.J., Schneeweiss, S. (2006a). 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**(5), 437–447.
- Stürmer, T., Rothman, K.J., Glynn, R.J. (2006b). Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiology and Drug Safety*, **15**(10), 698–709.
- Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, **101**(476), 1619–1637.
- Tan, Z. (2007). Understanding OR, PS, and DR. Comment on: “Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data”, by Kang, J.D.Y. and Schafer, J.L. *Statistical Science*, **22**(4), 560–568.
- Tan, Z. (2008). Improved local efficiency and double-robustness. Comment on: “Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis”, by Rubin, D.B. and van der Laan, M.J., *International Journal of Biostatistics*, **4**(1), Article 10.