Research Article

Statistics in Medicine

Received 25 June 2013, Accepted 23 April 2014 Published online 14 May 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6207

On regression adjustment for the propensity score

S. Vansteelandt[a*†] and R. M. Daniel[b]

Propensity scores are widely adopted in observational research because they enable adjustment for high-dimensional confounders without requiring models for their association with the outcome of interest. The results of statistical analyses based on stratification, matching or inverse weighting by the propensity score are therefore less susceptible to model extrapolation than those based solely on outcome regression models. This is attractive because extrapolation in outcome regression models may be alarming, yet difficult to diagnose, when the exposed and unexposed individuals have very different covariate distributions.

Standard regression adjustment for the propensity score forms an alternative to the aforementioned propensity score methods, but the benefits of this are less clear because it still involves modelling the outcome in addition to the propensity score. In this article, we develop novel insights into the properties of this adjustment method. We demonstrate that standard tests of the null hypothesis of no exposure effect (based on robust variance estimators), as well as particular standardised effects obtained from such adjusted regression models, are robust against misspecification of the outcome model when a propensity score model is correctly specified; they are thus not vulnerable to the aforementioned problem of extrapolation. We moreover propose efficient estimators for these standardised effects, which retain a useful causal interpretation even when the propensity score model is misspecified, provided the outcome regression model is correctly specified.

Keywords: causal inference; effect measure; model misspecification; positivity; pooling; propensity score; standardisation; strong confounding

1. Introduction

In a seminal paper in the 1980s, Rosenbaum and Rubin introduced the concept of propensity scores and laid out their key properties [1]. For a given exposure and baseline covariate set, they defined the propensity score for a given individual as the probability for that individual to be exposed, on the basis of his/her covariates. They subsequently demonstrated that when the covariate set is sufficient to control for confounding of the effect of exposure on outcome, then adjustment for the propensity score is also sufficient. Because the propensity score is scalar, this enables the use of simple and transparent methods for confounding adjustment that would otherwise be impossible in the presence of high-dimensional confounders [1, 2]: stratification by the propensity score [3, 4], matching on the propensity score [5] and inverse weighting by the propensity score [6, 7]. Each of these methods relies on a model for the propensity score, for example, a logistic regression model for a dichotomous exposure. They all avoid reliance on outcome regression models either by making raw comparisons within strata of individuals who are similar in terms of the propensity score or by making raw comparisons within the entire sample after reweighting the data based on the propensity score to account for confounding bias. These propensity score approaches are therefore less vulnerable to extrapolation than standard outcome regression approaches to adjust for confounding [4, 8]. Such extrapolation in outcome models is likely to occur when the exposed and unexposed individuals are very different in their confounder distributions. In that case, empirical information about the exposure effect is largely lacking, so that estimation under standard

[a] *Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent, Belgium*
[b] *Centre for Statistical Methodology and Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, U.K.*
**Correspondence to: S. Vansteelandt, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281 (S9), 9000 Gent, Belgium.*
[†] *E-mail: stijn.vansteelandt@ugent.be*

outcome regression approaches will almost completely rely on extrapolation of the covariate-specific outcome mean for the exposed to the observed covariate levels for the unexposed (and vice versa). This is worrisome as misspecification of the outcome model is probable and difficult to diagnose in such settings. Corresponding bootstrap or normal theory standard error estimators do not express the ensuing uncertainty because of model extrapolation.

Misspecification of the propensity score model is likewise a concern, but because propensity scores merely quantify the frequency of exposure at each covariate level, they do not enable transporting estimated associations between exposure and outcome from the unexposed to the exposed subpopulations (or vice versa) [8]. Lack of empirical information about the exposure effect therefore gets more explicitly (and thus more honestly) expressed in empty propensity-score-strata, lack of good matches, highly variable inverse probability weights and more generally, propensity-score-based exposure effect estimates with large standard errors.

The key property of propensity scores as summarising all confounders also justifies regression adjustment for the propensity score as a valid strategy for confounding adjustment [1, 2]. However, the benefits of this method are less evident because it still involves modelling the outcome. This may be the predominant reason why regression adjustment for the propensity score has not become as popular as the aforementioned propensity score methods, in spite of its simplicity. In this article, we develop novel insights into this adjustment method. In particular, we demonstrate that standard tests of the null hypothesis of no exposure effect (based on robust variance estimators), as well as particular standardised effects obtained from such adjusted regression models, are not vulnerable to the aforementioned extrapolation bias of outcome-based regression approaches. Our results shed new light on the problem of how to infer causal effects in the presence of strong confounding.

2. Regression adjustment for the propensity score

Let A denote a dichotomous exposure of interest, Y an outcome and C a set of baseline covariates. For example, A may denote right heart catheterization, and Y 30-day mortality. To address causal questions, we use the counterfactual framework [9, 10]. We let Y_a denote the value of the outcome that would have been observed had the exposure A been set, possibly contrary to fact, to level a. The propensity score [1] is defined as the probability of receiving the exposure or treatment, given the baseline covariates C, which we denote as $p(C) \equiv P(A = 1|C)$. It is usually obtained as the fitted value from a logistic regression model for the exposure, which we will refer to as the propensity score model.

Throughout, we assume the technical consistency assumption, which is generally presupposed in the causal inference literature and described elsewhere [11]. We moreover assume that treatment assignment is ignorable [9] (i.e. that the covariate set C is sufficient to adjust for confounding of the effect of A on Y), which can be technically formalised as that for all a,

$$Y_a \perp\!\!\!\perp A|C, \tag{1}$$

where $V \perp\!\!\!\perp W|Z$ for arbitrary random variables V, W and Z denotes that V is conditionally independent of W, given Z. The central result of Rosenbaum and Rubin [1] is then that also the propensity score is sufficient to adjust for confounding in the sense that

$$Y_a \perp\!\!\!\perp A|p(C). \tag{2}$$

This result justifies adjustment for confounding because of measured covariates C by including the propensity score along with the exposure in a model for the outcome. The resulting approach is referred to as regression adjustment for the propensity score [2] and will be the focus of this article.

Consider first linear regression adjustment for the propensity score. In particular, consider ordinary least squares estimation in model

$$E\{Y|a, p(c)\} = \theta_0 + \theta_1 a + \theta_2 p(c), \tag{3}$$

where $E(V|w)$ for arbitrary random variables V and W is shorthand notation for $E(V|W = w)$. By the ignorability assumption (2), θ_1 in model (3) equals the conditional exposure effect $E\left\{Y_1 - Y_0|p(c)\right\}$ if the outcome model (3) is correctly specified. In that case, the ordinary least squares estimator of θ_1 is an unbiased estimator of the conditional exposure effect $E\left\{Y_1 - Y_0|p(c)\right\}$ as well as the population-averaged

effect $E\left(Y_1 - Y_0\right)$. This continues to hold in large samples when the propensity score $p(c)$ in model (3) is replaced by the fitted value $\hat{p}(c)$ under a correctly specified propensity score model. Suppose however that the outcome model (3) is misspecified (e.g. that the dependence of the outcome on the propensity score is in fact nonlinear). Then, interestingly, the ordinary least squares estimator of θ_1 in model (3) (with $\hat{p}(c)$ in lieu of $p(c)$) continues to be an unbiased estimator of the population-averaged effect $E\left(Y_1 - Y_0\right)$ in large samples when the propensity score model is correctly specified and the exposure effect is homogenous in the sense that $E\left\{Y_1 - Y_0|p(c)\right\}$ is the same at all levels of the propensity score (see Robins, Mark and Newey [12]; see also the derivations in Appendix A.1). Under these conditions, the ordinary least squares estimator of θ_1 in model (3) does not necessitate correct specification of model (3) and is thus not vulnerable to extrapolation under that model.

Unfortunately, the previous results do not extend to nonlinear models. For instance, in the logistic regression model for a binary outcome,

$$P\left\{Y = 1|a, p(c)\right\} = \text{expit}\left\{\theta_0 + \theta_1 a + \theta_2 p(c)\right\}, \tag{4}$$

we still have, by analogy, that $\exp(\theta_1)$ equals the conditional exposure effect

$$\frac{P\left\{Y_1 = 1|p(c)\right\} P\left\{Y_0 = 0|p(c)\right\}}{P\left\{Y_1 = 0|p(c)\right\} P\left\{Y_0 = 1|p(c)\right\}}. \tag{5}$$

However, because of noncollapsibility of the odds ratio [13, 14], this generally differs from the conditional exposure effect

$$\frac{P\left(Y_1 = 1|c\right) P\left(Y_0 = 0|c\right)}{P\left(Y_1 = 0|c\right) P\left(Y_0 = 1|c\right)},$$

which is the focus of standard regression adjustment approaches. Moreover, it generally differs from the population-averaged exposure effect

$$\frac{P\left(Y_1 = 1\right) P\left(Y_0 = 0\right)}{P\left(Y_1 = 0\right) P\left(Y_0 = 1\right)},$$

which is the focus of certain propensity score approaches based on stratification, matching or inverse weighting. Furthermore, when the outcome model (4) is misspecified, the default maximum likelihood estimator of θ_1 in model (4) is generally a biased estimator of the conditional exposure effect (5), even when the propensity score model is known. This raises doubts over the usefulness of regression adjustment for the propensity score in nonlinear models, as it appears to demand correct specification of two models – an outcome model as well as a propensity score model – instead of one. In the following sections, we will develop novel insights into this adjustment method. In particular, we will demonstrate that the aforementioned properties of ordinary least squares adjustment for the propensity score carry over to testing of the null hypothesis of no exposure effects and to the estimation of particular standardised effect measures in nonlinear models.

3. Testing the causal null hypothesis

We will first consider testing of the causal null hypothesis that $E\left(Y_1 - Y_0|c\right) = 0$ for all covariate subgroups c. Consider in particular the logistic regression model for the outcome Y given by

$$P\left(Y = 1|a, c\right) = \text{expit}\left(\beta_0 + \beta_1 a + \beta_2' c\right).$$

Then, a test of the causal null hypothesis that $E\left(Y_1 - Y_0|c\right) = 0$ for all covariate subgroups c amounts to the default test that $\beta_1 = 0$. A limitation of this standard testing approach is that it may be severely liberal or conservative when this model is misspecified. This limitation is also shared by direct standardisation or so-called G-computation approaches [14–16] based on this model. These estimate the population-averaged effect $E\left(Y_1 - Y_0\right)$ as

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\text{expit}\left(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2'C_i\right)-\text{expit}\left(\hat{\beta}_0+\hat{\beta}_2'C_i\right)\right\},$$

where $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ are the maximum likelihood estimates of β_0, β_1 and β_2, and next evaluate whether this differs systematically from zero.

In Appendix A.2, we show that this limitation of traditional regression approaches is not shared by standard testing approaches that employ regression adjustment for the propensity score in a canonical generalised linear model (e.g. linear, loglinear and logistic models), for example, by standard tests of $\theta_1 = 0$ in model (4). In particular, we show that tests of the null hypothesis of no exposure effect in such generalised linear models that adjust for the propensity score and are based on a robust variance estimator do not falsely reject the null hypothesis more than $\alpha 100\%$ of the time in large samples (with α the nominal significance level) when the propensity score model is correctly specified, even when the outcome regression model is misspecified. They may however be mildly conservative when the outcome model is misspecified. This can be remedied by adjusting for the uncertainty in the estimated propensity score, as shown in Appendix A.2.

If the propensity score model is misspecified, then in general, adjustment for the estimated propensity score will no longer suffice to control for confounding. This can be remedied by including the confounders along with the propensity score in the outcome model, for example, by fitting model

$$P(Y=1|a,c)=\text{expit}\left\{\theta_0+\theta_1 a+\theta_2 p(c)+\theta_3'c\right\}.$$

Indeed, when the outcome model is correctly specified in the sense that $P(Y=1|a,c) = \text{expit}\left(\theta_0 + \theta_1 a + \theta_3' c\right)$, then $\theta_2 = 0$, and thus, the choice of propensity score model does not affect the validity of tests of the causal null hypothesis. In Appendix A.2, we show that the inclusion of covariates in the model, alongside the propensity score, only minimally affects the precision of the estimator of θ_1 when in truth $\theta_1 = 0$ and the propensity score model is correctly specified. The inclusion of additional covariates may however negatively impact precision when $\theta_1 \neq 0$ [17]; this is partly related to the fact that the adjustment also changes the meaning and magnitude of the exposure effect (Section 2). Simulation studies in Section 6 will investigate to what extent this adjustment also affects the precision of population-averaged effects $E\left(Y_1 - Y_0\right)$ estimated as

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\text{expit}\left(\hat{\theta}_0+\hat{\theta}_1+\hat{\theta}_2 p(C_i)+\hat{\theta}_3'C_i\right)-\text{expit}\left(\hat{\theta}_0+\hat{\theta}_2 p(C_i)+\hat{\theta}_3'C_i\right)\right\}, \quad (6)$$

where $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2$ and $\hat{\theta}_3$ are the maximum likelihood estimates of $\theta_0, \theta_1, \theta_2$ and θ_3.

4. Optimally weighted average exposure effects

The results of the previous section demonstrate that regression adjustment for the propensity score insulates tests of the causal null hypothesis against misspecification of the outcome regression model, provided that the propensity score is correctly modelled. However, this leaves unanswered the question of whether this robustness extends to certain effect size measures. As previously suggested, it does not extend to the exposure effect in the propensity score adjusted regression model itself, for example, to the parameter θ_1 in model (4). Likewise, it does not extend to the directly standardised exposure effect obtained from such regression model, for example, to the population-averaged effect (6). In this section, we demonstrate that it does extend to certain otherwise standardised causal effect measures on the difference and ratio scale.

Consider the weighted exposure effect

$$E\left\{w(C)(Y_1 - Y_0)\right\}, \quad (7)$$

where $w(C)$ is the standardised weight

$$w(C)=\frac{p(C)\{1-p(C)\}}{E\left[p(C)\{1-p(C)\}\right]}.$$

This expresses the effect that would be realised in a randomised experiment in which participating individuals were randomly selected from the population with probability proportional to $p(C)\{1-p(C)\}$. Individuals who, on the basis of their covariate data, are equally likely to be administered treatment or control in standard practice are thus most likely included in the study. Individuals who, on the basis of their covariate data, are unlikely to be administered experimental or control treatment are unlikely to be included in the study. Although in practice, such formal selection probabilities are not used to recruit participants into an experiment (e.g. a clinical trial), note that the low and high propensity strata are composed of individuals for whom most physicians would regard either experimental or control treatment inappropriate, and such individuals would only rarely be included in an experiment [18]. In observational studies with strong confounding, the weighted exposure effect (7) may therefore succeed better than the population-averaged effect $E\left(Y_1 - Y_0\right)$ to reflect the effect that would be attained in a typical clinical trial population.

The weight $w(C)$ can be viewed as being proportional to the amount of overlap between exposed and unexposed subjects. The data will thus in particular carry more information about the weighted effect (7) than about the population-averaged effect $E(Y_1 - Y_0)$ because information about the latter is weak when there is little overlap between exposed and unexposed individuals. This is formalised in Theorem 5.4 and Corollary 5.2 by Crump *et al*. [19], who show that of all estimands of the form $E\left\{w(C)(Y_1 - Y_0)\right\}$ for non-negative functions $w(C)$, the optimal one (in the sense of having the smallest semi-parametric efficiency bound) is the one considered earlier, whenever the conditional outcome variance (given A and C) is constant. Crump *et al*. [19] thus refer to the estimand (7) as an optimally weighted average treatment effect.

When the outcome is dichotomous, the optimally weighted average exposure effect (7) can easily be estimated by fitting the propensity score adjusted model (4) and then calculating

$$\sum_{i=1}^{n} \frac{\hat{p}(C_i)\left\{1-\hat{p}(C_i)\right\}}{\sum_{i=1}^{n}\hat{p}(C_i)\left\{1-\hat{p}(C_i)\right\}} \left\{\text{expit}\left(\hat{\theta}_0 + \hat{\theta}_1 + \hat{\theta}_2\hat{p}(C_i)\right) - \text{expit}\left(\hat{\theta}_0 + \hat{\theta}_2\hat{p}(C_i)\right)\right\}. \tag{8}$$

In Appendix A.3, we show that the resulting estimator has the attractive property of being unbiased in large samples when the propensity score model is correctly specified, even if the outcome model (4) is incorrect. This property extends to arbitrary canonical generalised linear models for the outcome that minimally include the exposure and propensity score, possibly along with other covariates, but not to other choices of weights in (7). We conclude that the estimator (8) shields the user from making extrapolation under the outcome regression model (4). It does that in two ways: once via its robustness against misspecification of the outcome model and once via the choice of weights $w(C)$, which accentuate parts of the data space where there is more information about the exposure effect.

Consider now the weighted exposure effect on the ratio scale:

$$\frac{E\left\{w(C)Y_1\right\}}{E\left\{w(C)Y_0\right\}}. \tag{9}$$

When the outcome is dichotomous, this effect can be estimated by fitting the propensity score adjusted model (4) and then calculating

$$\frac{\sum_{i=1}^{n}\hat{p}(C_i)\left\{1-\hat{p}(C_i)\right\}\text{expit}\left(\hat{\theta}_0 + \hat{\theta}_1 + \hat{\theta}_2\hat{p}(C_i)\right)}{\sum_{i=1}^{n}\hat{p}(C_i)\left\{1-\hat{p}(C_i)\right\}\text{expit}\left(\hat{\theta}_0 + \hat{\theta}_2\hat{p}(C_i)\right)}.$$

This estimator is no longer guaranteed to be unbiased when the outcome model (4) is misspecified, even when the propensity score model is correct. In Appendix A.4, we show that this can be remedied by including an exposure by propensity score interaction in the outcome model. In particular, consider the propensity score adjusted model

$$P\{Y=1|a,p(c)\} = \text{expit}\left\{\theta_0 + \theta_1 a + \theta_2 p(c) + \theta_3 a p(c)\right\}, \tag{10}$$

and let $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2$ and $\hat{\theta}_3$ be the maximum likelihood estimates of $\theta_0, \theta_1, \theta_2$ and θ_3 in this model with $p(c)$ substituted by $\hat{p}(c)$. It then follows from the results in Appendix A.4 that the estimator

$$\frac{\sum_{i=1}^{n} \hat{p}(C_i)\left\{1-\hat{p}(C_i)\right\} \operatorname{expit}\left(\hat{\theta}_0+\hat{\theta}_1+(\hat{\theta}_2+\hat{\theta}_3)\hat{p}(C_i)\right)}{\sum_{i=1}^{n} \hat{p}(C_i)\left\{1-\hat{p}(C_i)\right\} \operatorname{expit}\left(\hat{\theta}_0+\hat{\theta}_2\hat{p}(C_i)\right)} \tag{11}$$

is unbiased in large samples when the propensity score model is correct, even if the outcome model (4) is incorrect.

Confidence intervals for estimators (8) and (11) can be based on the bootstrap or sandwich standard error estimators (see the Web Appendix for details). In the following section, we study the efficiency of these estimators.

5. Efficient estimation of (optimally) weighted average exposure effects

It can be shown that, up to asymptotic equivalence, all estimators that are consistent for the optimally weighted average exposure effect (7) when the propensity score model is correctly specified can be obtained as the solution to

$$0=\sum_{i=1}^{n}\left\{A_i-\hat{p}(C_i)\right\}\left\{Y_i-\phi(C_i)\right\}-\psi\hat{p}(C_i)\left\{1-\hat{p}(C_i)\right\} \tag{12}$$

for some scalar function $\phi(C)$. For instance, the estimator given in (8) is obtained by choosing

$$\phi(C)=\{1-\hat{p}(C)\}\operatorname{expit}\left(\hat{\theta}_0+\hat{\theta}_1+\hat{\theta}_2 p(C)\right)+\hat{p}(C)\operatorname{expit}\left(\hat{\theta}_0+\hat{\theta}_2 p(C)\right); \tag{13}$$

this can be verified upon noting that $\sum_{i=1}^{n}\left\{A_i-\hat{p}(C_i)\right\}Y_i = \sum_{i=1}^{n}\left\{A_i-\hat{p}(C_i)\right\}\operatorname{expit}\left(\hat{\theta}_0+\hat{\theta}_1 A_i +\hat{\theta}_2\hat{p}(C_i)\right)$. The efficient estimator within this class of estimators that solves (12) for arbitrary $\phi(C)$ is obtained upon choosing

$$\phi(C)=\{1-p(C)\}E(Y|A=1,C)+p(C)E(Y|A=0,C). \tag{14}$$

This can be seen upon noting that this expression for $\phi(C)$ equals the population least squares projection of $\{A-p(C)\}Y-\psi p(C)\{1-p(C)\}$ onto $A-p(C)$ (given C).

Calculating the efficient score requires substituting $p(C)$ and $E(Y|A,C)$ in (14) by their fitted values $\hat{p}(C)$ and $\hat{E}(Y|A,C)$ under suitable models. The efficiency is therefore local: it is only attained under a correctly specified propensity score model, when also the model for the outcome mean $E(Y|A,C)$ is correct. Upon plugging the expression (14) for $\phi(C)$ into (12), it is seen that the efficient estimator of the optimally weighted average exposure effect (7) equals

$$\frac{\sum_{i=1}^{n}\left\{A_i-\hat{p}(C_i)\right\}\left\{Y_i-\hat{E}(Y|A_i,C_i)\right\}}{\sum_{i=1}^{n}\hat{p}(C_i)\left\{1-\hat{p}(C_i)\right\}}+\frac{\sum_{i=1}^{n}\left\{1-\hat{p}(C_i)\right\}\hat{p}(C_i)\left\{\hat{E}(Y|A_i=1,C_i)-\hat{E}(Y|A_i=0,C_i)\right\}}{\sum_{i=1}^{n}\hat{p}(C_i)\left\{1-\hat{p}(C_i)\right\}}. \tag{15}$$

An efficient estimator may alternatively be obtained by calculating (8) based on an outcome model that includes all covariates in addition to the exposure and propensity score; this can be understood upon contrasting (13) and (14).

By construction, the efficient estimator given in (15) is an unbiased estimator of the optimally weighted average exposure effect (7) (in large samples) when the propensity score model is correctly specified. When the propensity score model is misspecified, then as the sample size increases, the estimated propensity scores $\hat{p}(C)$ converge to some value $\tilde{p}(C)$ that may differ from the actual propensity score. When the outcome model for $E(Y|A,C)$ is correctly specified, then it is readily seen from expression (15) for the efficient estimator that it converges to

$$\frac{E\left[\tilde{p}(C)\{1-\tilde{p}(C)\}(Y_1-Y_0)\right]}{E\left[\tilde{p}(C)\{1-\tilde{p}(C)\}\right]}. \tag{16}$$

It thus continues to be an unbiased estimator (in large samples) of a weighted average exposure effect, although based on different weights than those given in (7). In particular, it retains a causal interpretation when the propensity score model is misspecified, provided that the outcome model is correctly specified; this is not the case for the estimator (8), by not relying on a separate outcome model.

A similar reasoning as in the previous paragraph shows that, up to asymptotic equivalence, all estimators that are consistent for the weighted average exposure effect $E\left\{w(C)Y_1\right\}/E\left\{w(C)Y_0\right\}$ when the propensity score model is correctly specified can be obtained as the solution to:

$$0=\sum_{i=1}^{n}A_i\left\{1-\hat{p}(C_i)\right\}Y_i-(1-A_i)\hat{p}(C_i)\psi Y_i-\phi(C_i)\left\{A_i-\hat{p}(C_i)\right\}$$

for some scalar function $\phi(C)$. The (locally) efficient estimator in this class is given by

$$\frac{\sum_{i=1}^{n}\left\{1-\hat{p}(C_i)\right\}A_i\left\{Y_i-\hat{E}(Y_i|A_i=1,C_i)\right\}+\left\{1-\hat{p}(C_i)\right\}\hat{p}(C_i)\hat{E}(Y_i|A_i=1,C_i)}{\sum_{i=1}^{n}\left(1-A_i\right)\hat{p}(C_i)\left\{Y_i-\hat{E}(Y_i|A_i=0,C_i)\right\}+\left\{1-\hat{p}(C_i)\right\}\hat{p}(C_i)\hat{E}(Y_i|A_i=0,C_i)}.$$

An efficient estimator may alternatively be obtained by calculating (11) based on an outcome model that includes all covariates in addition to the exposure, propensity score and their interaction. It is an unbiased estimator (in large samples) for the weighted average exposure effect

$$\frac{E\left[\tilde{p}(C)\left\{1-\tilde{p}(C)\right\}Y_1\right]}{E\left[\tilde{p}(C)\left\{1-\tilde{p}(C)\right\}Y_0\right]}$$

when the propensity score model is correct (in which case, it coincides with (9)) or when the outcome model is correctly specified.

6. Simulation study

6.1. Design

In each simulation scenario, we simulate 1000 datasets, each of size n, following a data-generating mechanism that slightly generalises that in Stürmer *et al*. [20] by evaluating the impact of model misspecification. The simulation is carried out in two stages. In the first stage, six independent confounders C_1–C_6 are simulated: C_1–C_3 are Bernoulli random variables each with mean 0.2 and C_4–C_6 are each standard normal. The so-called intended treatment variable, T, is then simulated from a Bernoulli distribution with mean

$$P\left(T=1\left|C_1,\ldots,C_6\right.\right)=\text{expit}(\alpha_0+\alpha_1C_1+\cdots+\alpha_6C_6+\alpha_{16}C_1C_6).$$

Then, in the second stage, two further binary covariates C_7 and C_8 are generated. C_7 is most likely to be present (= 1) when the intended treatment T is least likely; it is set to 1 with probability $\left\{\gamma-P\left(T=1\left|C_1,\ldots,C_6\right.\right)\right\}$. C_8 is likely to be present (= 1) when the intended treatment T is most likely; it is set to 1 with probability $\left\{P\left(T=1\left|C_1,\ldots,C_6\right.\right)-\delta\right\}$. The values of γ and δ are chosen such that the prevalence of C_7 and C_8 is close to 0.1. The actual treatment A is then simulated from a Bernoulli distribution with mean

$$P\left(A=1\left|C_1,\ldots,C_8\right.\right)=\text{expit}(\alpha_0+\alpha_1C_1\cdots+\alpha_8C_8+\alpha_{16}C_1C_6).$$

Finally, given C_1–C_8 and A, the outcome Y is simulated from a Bernoulli distribution with mean

$$P\left(Y=1\left|C_1,\ldots,C_8,A\right.\right)=\text{expit}(\beta_0+\beta_1C_1\cdots+\beta_8C_8+\beta_{16}C_1C_6+\beta_AA);$$

β_0 is chosen such that the prevalence of Y in each scenario is about 0.1.

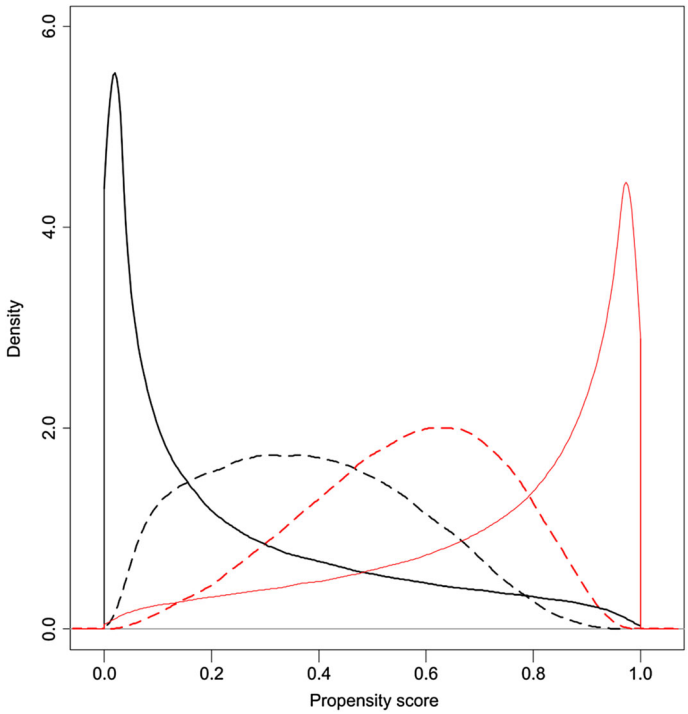

Figure 1. Simulation study: propensity score distribution for exposed (red) and unexposed (black) individuals; dashed for good overlap and solid for poor overlap.

Different scenarios are considered by varying n, the α's and the β's (see Tables 1 and 2 in the Supporting information). We look at the impact of good/poor overlap by varying the α's (Table 1 in the Supporting information, and Figure 1), we vary the magnitude of β_A, from null to weak to strong (Table 2 in the Supporting information) and we change n from 100 to 1000, to look at the effect of sample size.

A logistic regression model for A given C_1–C_8 is fitted for the propensity score. This model is either correctly specified by including all main effects of C_1–C_8 along with a product term C_1C_6 or incorrectly specified by omitting the product term. A logistic regression model is also fitted for the outcome Y, with A as an independent variable, in addition to C_1–C_8 ('Cov-adj'), the propensity score ("PS-adj") or both C_1–C_8 and the propensity score ('Cov-and-PS-adj'). When adjusting for C_1–C_8, the product term C_1C_6 is either included or omitted, to look at the effect of incorrectly specifying this model. When estimating a risk ratio, the interaction between A and the propensity score is additionally included when adjusting for the propensity score. From each of these outcome models, we predict Y_1 and Y_0 for each subject, which are then averaged, and four estimators formed: the marginal (population-averaged) risk difference/ratio and the (optimally) weighted risk difference/ratio, as defined in Section 4.

We also consider inverse probability weighted estimators obtained by fitting a logistic regression model for the outcome that includes only the exposure ('IPW') or the exposure along with covariates C_1–C_8 and the product term C_1C_6 ('Cov-adj IPW'). In both cases, we weigh each subject i's contribution to the estimation by $A_i/\hat{p}(C_i) + (1 - A_i)/\{1 - \hat{p}(C_i)\}$ based on either the correctly specified or incorrectly specified propensity score model. Because of the anticipated poor performance of the methods based on inverse weighting in the presence of extreme weights, we include two further estimators, ('IPW (Pet)' and 'Cov-adj IPW (Pet)') which are the same as 'IPW' and 'Cov-adj IPW', respectively, except that the weights are truncated at the 2.5th and 97.5th percentiles, as suggested by Petersen *et al*. [21].

6.2. *Results*

The results are displayed in Tables I–III, Figures 2–4, together with Tables 3–9 in the Supporting information, and confirm what the theory set out in this paper suggests, as we now discuss.

6.2.1. Testing the causal null hypothesis. Evaluation of type I error rates and power is based on Wald tests using robust standard errors, ignoring that the propensity score is estimated. We find that adjustment for the propensity score leads to valid tests of the null hypothesis (Table I). Looking at the first and fourth

Table I. Simulation results: tests of the null hypothesis.

			Both correct		PS incorrect		Cov incorrect		Both incorrect	
			Proportion of H_0 rejected at 5%							
Method	Overlap	Causal	$n = 100$	1000	100	1000	100	1000	100	1000
Cov-adj			0.063	0.051	0.063	0.051	0.062	0.197	0.062	0.197
PS-adj			0.044	0.049	0.044	0.182	0.044	0.049	0.044	0.182
Cov-and-PS-adj	Good	Null	0.066	0.052	0.068	0.051	0.068	0.052	0.074	0.182
IPW			0.050	0.046	0.047	0.193	0.050	0.046	0.047	0.193
Cov-adj IPW			0.093	0.054	0.077	0.050	0.075	0.057	0.074	0.212
Cov-adj			0.067	0.341	0.067	0.341	0.085	0.774	0.085	0.774
PS-adj			0.034	0.312	0.052	0.756	0.034	0.312	0.052	0.756
Cov-and-PS-adj	Good	Weak	0.075	0.340	0.064	0.340	0.072	0.337	0.079	0.752
IPW			0.050	0.290	0.078	0.746	0.050	0.290	0.078	0.746
Cov-adj IPW			0.082	0.338	0.074	0.349	0.080	0.333	0.110	0.782
Cov-adj			0.151	0.987	0.151	0.987	0.223	1.000	0.223	1.000
PS-adj			0.101	0.979	0.176	1.000	0.101	0.979	0.176	1.000
Cov-and-PS-adj	Good	Strong	0.153	0.986	0.153	0.985	0.143	0.984	0.220	1.000
IPW			0.148	0.951	0.205	0.998	0.148	0.951	0.205	0.998
Cov-adj IPW			0.179	0.978	0.167	0.980	0.189	0.970	0.248	1.000
Cov-adj			0.056	0.050	0.056	0.050	0.064	0.127	0.064	0.127
PS-adj			0.058	0.035	0.054	0.087	0.058	0.035	0.054	0.087
Cov-and-PS-adj	Poor	Null	0.083	0.043	0.081	0.039	0.085	0.045	0.088	0.106
IPW			0.104	0.106	0.129	0.228	0.104	0.106	0.129	0.228
Cov-adj IPW			0.142	0.109	0.125	0.100	0.142	0.116	0.153	0.256
Cov-adj			0.073	0.292	0.073	0.292	0.080	0.592	0.080	0.592
PS-adj			0.063	0.242	0.068	0.464	0.063	0.242	0.068	0.464
Cov-and-PS-adj	Poor	Weak	0.095	0.264	0.089	0.273	0.086	0.254	0.103	0.485
IPW			0.133	0.329	0.159	0.566	0.133	0.329	0.159	0.566
Cov-adj IPW			0.155	0.302	0.139	0.307	0.166	0.314	0.182	0.563
Cov-adj			0.126	0.923	0.126	0.923	0.160	0.989	0.160	0.989
PS-adj			0.085	0.868	0.128	0.978	0.085	0.868	0.128	0.978
Cov-and-PS-adj	Poor	Strong	0.125	0.896	0.124	0.900	0.118	0.885	0.150	0.972
IPW			0.239	0.753	0.294	0.883	0.239	0.753	0.294	0.883
Cov-adj IPW			0.204	0.783	0.195	0.787	0.223	0.772	0.249	0.910

'Both correct' indicates that both the model for the propensity score and for the outcome, if used, were correctly specified. 'PS incorrect' refers to the situation in which only the propensity score model is misspecified, 'Cov incorrect' to the situation in which only the outcome model is misspecified and 'Both incorrect' to the situation in which both models are misspecified.

sixths of this table (where the causal effect is null), we would expect all methods, in the absence of model misspecification, to give tests that reject the null hypothesis in around 5% of simulated datasets. Note, however, that when the overlap is poor, the weights in the inverse weighting approach become extreme so that the two tests based on inverse weighting have a poor performance, even with a sample size of 1000. Propensity score adjustment, on the other hand, does not suffer from this problem. Using the truncated weights suggested by Petersen *et al*. (see Table 5 in the Supporting information) improves the performance of inverse weighting slightly but not fully; this is likely due to a bias–variance trade-off discussed later.

Because we ignored the estimation of the propensity score in the estimation of the standard errors in which these tests were based, we would expect the estimated standard errors to be slightly too large and thus that the tests would be slightly conservative. This is seen in most but not quite all settings that rely on the propensity score. For example, for the small sample size and when the overlap is poor, propensity score adjustment rejects the null in 5.8% of simulations. Closer inspection (not reported here) reveals

Table II. Simulation results: estimation of risk differences (given as percentages), $n = 100$.

					Both correct		PS incorrect		Cov incorrect		Both incorrect	
Estimand	Estimator	Overlap	Causal effect	True value	Bias	Actual SE	Bias	Actual SE	Bias	Actual SE	Bias	Actual SE
	Cov-adj				−0.476	7.751	−0.476	7.751	1.682	7.437	1.682	7.437
	PS-adj				−0.342	6.679	1.673	6.406	−0.342	6.679	1.673	6.406
Marginal	Cov-and-PS-adj	Good	Null	0.000	−0.559	7.818	−0.502	7.626	−0.523	7.794	1.394	7.315
	IPW				−0.108	7.136	1.912	6.658	−0.108	7.136	1.912	6.658
	Cov-adj IPW				−0.602	8.211	−0.581	7.980	0.042	7.913	1.651	7.649
	Cov-adj				−0.473	7.714	−0.495	7.803	1.655	7.433	1.670	7.428
Opt-W	PS-adj	Good	Null	0.000	−0.315	6.641	1.669	6.418	−0.315	6.641	1.669	6.418
	Cov-and-PS-adj				−0.576	7.838	−0.521	7.717	−0.523	7.789	1.398	7.415
	Cov-adj				−0.280	7.662	−0.280	7.662	1.750	7.381	1.750	7.381
	PS-adj				−0.627	6.420	1.354	6.294	−0.627	6.420	1.354	6.294
Marginal	Cov-and-PS-adj	Good	Weak	3.117	−0.490	7.609	−0.521	7.146	−0.445	7.544	1.288	7.087
	IPW				−0.402	6.582	1.360	6.417	−0.402	6.582	1.360	6.417
	Cov-adj IPW				−0.383	7.878	−0.354	7.765	0.104	7.763	1.660	7.593
	Cov-adj			3.201	−0.400	7.610	−0.461	7.685	1.634	7.352	1.575	7.352
Opt-W	PS-adj	Good	Weak	or	−0.690	6.390	1.284	6.291	−0.756	6.390	1.218	6.291
	Cov-and-PS-adj			3.266	−0.580	7.603	−0.608	7.296	−0.537	7.527	1.203	7.120
	Cov-adj				−0.292	7.481	−0.292	7.481	1.585	7.619	1.585	7.619
	PS-adj				−1.291	6.445	0.481	6.652	−1.291	6.445	0.481	6.652
Marginal	Cov-and-PS-adj	Good	Strong	7.777	−0.665	7.408	−0.914	7.230	−0.821	7.145	0.723	7.405
	IPW				−1.357	6.865	0.398	6.864	−1.357	6.865	0.398	6.864
	Cov-adj IPW				−0.566	7.704	−0.463	7.583	−0.063	7.590	1.422	7.790
	Cov-adj			7.983	−0.591	7.395	−0.684	7.485	1.307	7.561	1.147	7.557
Opt-W	PS-adj	Good	Strong	or	−1.506	6.457	0.309	6.666	−1.661	6.457	0.154	6.666
	Cov-and-PS-adj			8.138	−0.839	7.366	−1.163	7.291	−1.018	7.122	0.467	7.449
	Cov-adj				−0.568	9.591	−0.568	9.591	1.373	9.684	1.373	9.684
	PS-adj				−0.690	9.409	1.182	9.182	−0.690	9.409	1.182	9.182
Marginal	Cov-and-PS-adj	Poor	Null	0.000	−0.674	10.903	−0.716	10.495	−0.703	10.524	1.039	10.648
	IPW				1.136	10.866	2.885	11.008	1.136	10.866	2.885	11.008
	Cov-adj IPW				−0.397	11.583	−0.379	11.272	0.199	11.738	1.487	11.550
	Cov-adj				−0.563	9.523	−0.560	9.557	1.342	9.510	1.375	9.458
Opt-W	PS-adj	Poor	Null	0.000	−0.498	8.948	1.265	8.787	−0.498	8.948	1.265	8.787
	Cov-and-PS-adj				−0.546	10.596	−0.648	10.374	−0.646	10.180	1.085	10.313
	Cov-adj				−0.407	9.386	−0.407	9.386	1.646	9.422	1.646	9.422
	PS-adj				−1.040	9.219	0.653	9.051	−1.040	9.219	0.653	9.051
Marginal	Cov-and-PS-adj	Poor	Weak	3.149	−0.651	11.033	−0.670	10.840	−0.641	10.736	1.136	10.971
	IPW				0.214	10.267	1.935	10.445	0.214	10.267	1.935	10.445
	Cov-adj IPW				−1.020	11.206	−0.959	10.914	−0.183	11.307	1.011	11.237
	Cov-adj			3.324	−0.504	9.319	−0.531	9.347	1.432	9.211	1.407	9.176
Opt-W	PS-adj	Poor	Weak	or	−1.092	8.721	0.514	8.591	−1.117	8.721	0.489	8.591
	Cov-and-PS-adj			3.349	−0.763	10.772	−0.790	10.630	−0.780	10.333	0.934	10.524
	Cov-adj				−1.380	9.425	−1.380	9.425	0.592	9.825	0.592	9.825
	PS-adj				−2.175	8.733	−0.674	8.892	−2.175	8.733	−0.674	8.892
Marginal	Cov-and-PS-adj	Poor	Strong	7.849	−1.801	10.802	−1.657	10.315	−1.762	10.659	0.002	10.740
	IPW				−0.957	10.186	0.461	9.836	−0.957	10.186	0.461	9.836
	Cov-adj IPW				−1.647	11.040	−1.544	10.687	−1.007	11.217	0.139	11.225
	Cov-adj			8.286	−1.750	9.457	−1.751	9.505	0.025	9.736	0.002	9.762
Opt-W	PS-adj	Poor	Strong	or	−2.583	8.365	−1.172	8.567	−2.641	8.365	−1.230	8.567
	Cov-and-PS-adj			8.344	−2.029	10.480	−1.975	10.258	−2.130	10.157	−0.528	10.469

'Opt-W' stands for 'optimally weighted'. The description of 'Both correct', 'PS incorrect', 'Cov incorrect' and 'Both incorrect' are as given in the caption of Table I. When the propensity score model is misspecified, the weights used to define the optimally weighted estimand are different, leading to a different true value (except under the null); the first true value given is for the setting in which the propensity score model is correctly specified, and the alternative applies when the propensity score model is misspecified.

Table III. Simulation results: estimation of risk differences (given as percentages), $n = 1000$.

					Both correct		PS incorrect		Cov incorrect		Both incorrect	
Estimand	Estimator	Overlap	Causal effect	True value	Bias	Actual SE	Bias	Actual SE	Bias	Actual SE	Bias	Actual SE
	Cov-adj				0.056	1.981	0.056	1.981	2.189	1.921	2.189	1.921
	PS-adj				0.051	1.996	2.236	1.948	0.051	1.996	2.236	1.948
Marginal	Cov-and-PS-adj	Good	Null	0.000	0.054	1.983	0.052	1.984	0.056	1.988	2.080	1.937
	IPW				0.043	2.091	2.413	1.970	0.043	2.091	2.413	1.970
	Cov-adj IPW				0.047	2.052	0.046	2.014	0.071	2.067	2.318	1.935
	Cov-adj				0.060	2.031	0.058	2.073	2.282	2.001	2.262	1.983
Opt-W	PS-adj	Good	Null	0.000	0.053	2.026	2.271	1.976	0.053	2.026	2.271	1.976
	Cov-and-PS-adj				0.058	2.034	0.053	2.075	0.059	2.037	2.147	1.998
	Cov-adj				−0.060	1.881	−0.060	1.881	2.022	1.871	2.022	1.871
	PS-adj				−0.045	1.919	2.107	1.904	−0.045	1.919	2.107	1.904
Marginal	Cov-and-PS-adj	Good	Weak	3.117	−0.071	1.892	−0.062	1.887	−0.061	1.893	1.924	1.892
	IPW				−0.047	2.030	2.198	1.946	−0.047	2.030	2.198	1.946
	Cov-adj IPW				−0.043	1.981	−0.048	1.933	−0.014	2.003	2.140	1.904
	Cov-adj			3.201	−0.061	1.933	−0.064	1.969	2.170	1.953	2.054	1.934
Opt-W	PS-adj	Good	Weak	or	−0.080	1.949	2.111	1.931	−0.145	1.949	2.046	1.931
	Cov-and-PS-adj			3.266	−0.072	1.944	−0.066	1.974	−0.064	1.944	1.945	1.952
	Cov-adj				0.043	1.854	0.043	1.854	1.991	1.853	1.991	1.853
	PS-adj				0.105	1.907	2.130	1.904	0.105	1.907	2.130	1.904
Marginal	Cov-and-PS-adj	Good	Strong	7.777	0.043	1.868	0.053	1.865	0.056	1.872	1.916	1.882
	IPW				0.016	2.022	2.109	1.905	0.016	2.022	2.109	1.905
	Cov-adj IPW				0.034	1.935	0.044	1.888	0.065	1.952	2.038	1.863
	Cov-adj			7.983	0.035	1.914	0.041	1.933	2.206	1.924	1.960	1.910
Opt-W	PS-adj	Good	Strong	or	0.016	1.939	2.083	1.924	−0.139	1.939	1.928	1.924
	Cov-and-PS-adj			8.138	0.035	1.928	0.046	1.942	0.046	1.930	1.863	1.935
	Cov-adj				0.059	2.343	0.059	2.343	2.087	2.390	2.087	2.390
	PS-adj				0.038	2.598	1.859	2.585	0.038	2.598	1.859	2.585
Marginal	Cov-and-PS-adj	Poor	Null	0.000	0.015	2.442	0.016	2.427	0.032	2.506	1.738	2.520
	IPW				0.420	5.387	3.312	5.279	0.420	5.387	3.312	5.279
	Cov-adj IPW				0.041	3.555	0.065	3.375	0.280	3.657	2.855	3.598
	Cov-adj				0.077	2.458	0.077	2.487	2.158	2.464	2.163	2.470
Opt-W	PS-adj	Poor	Null	0.000	0.052	2.580	1.861	2.574	0.052	2.580	1.861	2.574
	Cov-and-PS-adj				0.030	2.562	0.030	2.577	0.045	2.579	1.806	2.602
	Cov-adj				0.091	2.473	0.091	2.473	2.189	2.493	2.189	2.493
	PS-adj				0.279	2.678	2.078	2.653	0.279	2.678	2.078	2.653
Marginal	Cov-and-PS-adj	Poor	Weak	3.149	0.076	2.529	0.060	2.519	0.143	2.588	1.841	2.594
	IPW				0.516	5.470	3.426	5.554	0.516	5.470	3.426	5.554
	Cov-adj IPW				0.024	3.671	0.028	3.493	0.263	3.794	2.779	3.808
	Cov-adj			3.324	0.094	2.625	0.104	2.650	2.181	2.593	2.170	2.601
Opt-W	PS-adj	Poor	Weak	or	0.101	2.683	1.887	2.659	0.075	2.683	1.862	2.659
	Cov-and-PS-adj			3.349	0.084	2.686	0.076	2.701	0.083	2.688	1.816	2.707
	Cov-adj				−0.065	2.220	−0.065	2.220	2.095	2.302	2.095	2.302
	PS-adj				0.319	2.446	2.068	2.481	0.319	2.446	2.068	2.481
Marginal	Cov-and-PS-adj	Poor	Strong	7.849	−0.095	2.321	−0.105	2.321	0.041	2.391	1.706	2.442
	IPW				0.402	3.731	2.813	4.268	0.402	3.731	2.813	4.268
	Cov-adj IPW				−0.046	3.123	−0.008	2.917	0.182	3.218	2.516	3.248
	Cov-adj			8.286	−0.120	2.383	−0.084	2.401	1.965	2.408	1.956	2.429
Opt-W	PS-adj	Poor	Strong	or	−0.149	2.484	1.597	2.514	−0.207	2.484	1.539	2.514
	Cov-and-PS-adj			8.344	−0.143	2.491	−0.119	2.508	−0.147	2.499	1.560	2.576

For a more detailed explanation, see the caption for Table II.

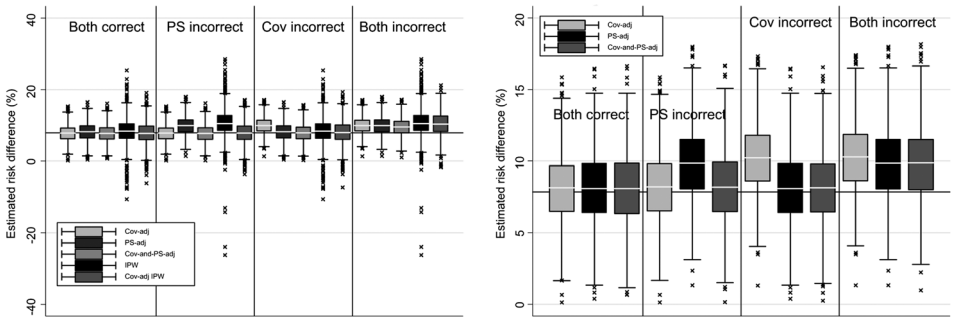

Figure 2. Box plots for the estimated risk differences over 1000 simulations. Left: estimators of the marginal effect; right: estimators of the optimally weighted average exposure effect. The simulation scenario depicted here is for $n = 1000$, poor overlap and strong causal effect. The black horizontal lines denote the true effect. The outer limits of each box denote the 25th and 75th percentiles, the white horizontal line inside the box denotes the median, the outer whiskers denote the lower and upper adjacent values (defined to be a distance 1.5×the interquartile range below and above the 25th and 75th percentile, respectively), and the crosses beyond the whiskers are considered to be 'outliers'. The description of 'Both correct', 'PS incorrect', 'Cov incorrect' and 'Both incorrect' are as given in the caption for Table I.

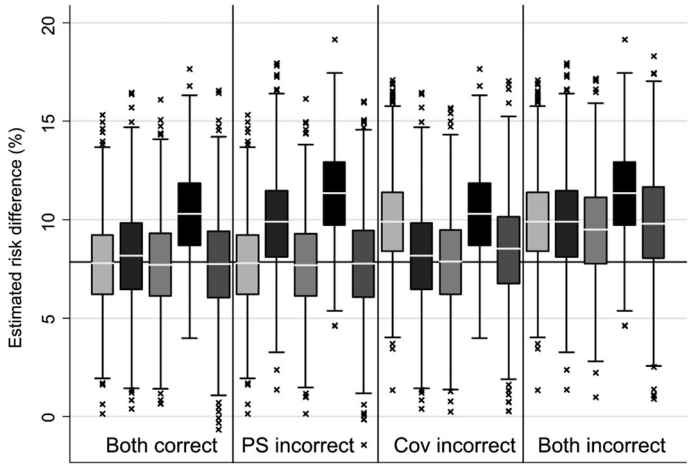

Figure 3. This is the same as the left-hand side of Figure 2, except that the Petersen *et al.* truncation has been applied to the inverse probability weights in the fourth and fifth estimators. The legend is the same as for the left-hand side of Figure 2.

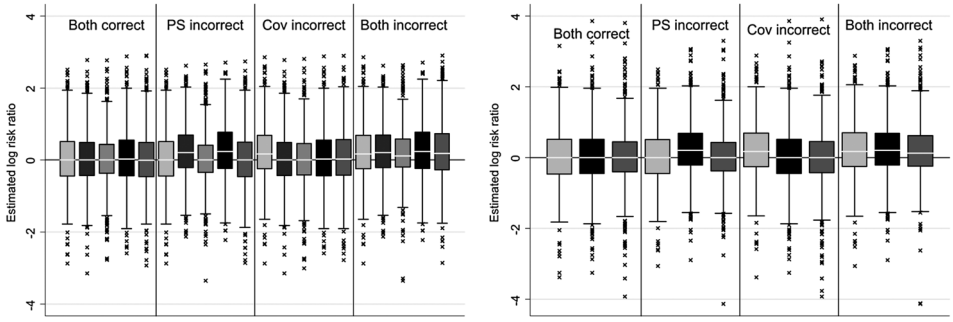

Figure 4. Box plots for the estimated log risk ratios over 1000 simulations. Left: estimators of the marginal effect; right: estimators of the weighted exposure effect. The simulation scenario depicted here is for $n = 100$, good overlap and null causal effect. The legends are the same as for those in Figure 2.

that the standard errors are indeed slightly overestimated even in this scenario; because the bias in the effect estimates is also negligible, we conclude that the slightly liberal nature of the test in this case is a consequence of the finite sample non-normality of the test statistic.

Table I also shows that tests of the null hypothesis based on both covariate and propensity score adjustment are robust to misspecification of one or the other of the two models. Away from the null, we see some increase in power from additionally adjusting for covariates in the outcome model, compared with only adjusting for the propensity score, even when the outcome model is misspecified; this is more pronounced at the lower sample size.

6.2.2. Estimating marginal and (optimally) weighted exposure effects. Figures 2 (showing results for the risk difference at the larger sample size, with poor overlap and a strong causal effect) and 4 (showing results for the log risk ratio at the smaller sample size, with good overlap and a null causal effect) are chosen as examples of the performance of the various methods in estimating the marginal and (optimally) weighted exposure effects. The graphs from all other scenarios showed a similar overall picture.

First, we note that the poor performance of the tests based on inverse weighting techniques when the overlap is poor carries through also to effect estimation. These estimators are inefficient (see the left-hand side of Figure 2) and exhibit more finite sample bias than other methods, even at a sample size of 1000 (this is more pronounced for the log risk ratio: see Tables 3 and 4 in the Supporting information). Figure 3 shows how the left-hand side of Figure 2 changes upon truncating the weights. We see that there is a marked decrease in the variability of the IPW estimates, as we would expect; however, this comes at the price of also introducing a marked (upward) bias.

Both Figures 2 and 4 demonstrate the robustness property of the estimator based on both covariate and propensity score adjustment for the (optimally) weighted average exposure effect. Only in the final quadrant of the graph does the median of the 1000 estimates deviate appreciably from the truth. This property appears to extend to the marginal effects, which is due to the lack of heterogeneity of the causal effect at different levels of the propensity score in these simulations. In other simulations (not presented here), with larger heterogeneity, the robustness against misspecification of the outcome model was (slightly) lost for the marginal effects, although remained (as theory would suggest) for the (optimally) weighted average exposure effects. In Figure 4, we see evidence of very slight gains in efficiency from additionally adjusting for the covariates in the outcome model, compared with only adjusting for the propensity score, even when the outcome model is misspecified. However, Tables II and III and Tables 3 and 4 in the Supporting information show that this is not universal over all scenarios and estimands.

7. Reanalysis of data on right heart catheterization

We consider a reanalysis of observational data on right heart catheterization in hospitalised adult patients at five medical centres in the U.S.A., to infer its effect on 30-day mortality [22]. Right heart catheterization is a diagnostic procedure used for critically ill patients. Detailed information about the study and the nature of the variables is given by Connors *et al*. [22], who found (based on propensity-score-matching) that right heart catheterization leads to lower survival rates.

Of the 5735 individuals in the study, 2184 received right heart catheterization within 24 h of admission. Of the latter, 38.0% died within 30 days from admission versus 30.6% of those who did not receive right heart catheterization within 24 h of admission. These results are difficult to interpret because both treatment groups differ substantially in many covariates available upon admission. As in [23], we therefore constructed a propensity score based on a logistic model that includes all 72 available baseline covariates. Logistic regression adjustment for the propensity score resulted in a conditional treatment effect odds ratio of 1.30 (95% confidence interval 1.14 to 1.48), corresponding with a population risk difference of 6.46% (SE 1.34%) based on direct standardisation (i.e. estimator (6)). The benefit of these analysis results relative to standard logistic regression adjustment is questionable because they demand correct specification of both the propensity score model and the logistic outcome model (except for testing the null hypothesis of no treatment effect that, by the results of Section 3, is valid when solely the propensity score is correctly specified). We therefore translated them to an optimally weighted average exposure effect of 6.48% (SE 1.34%), based on estimator (8). Using a logistic outcome model that includes all 72 covariates along with interactions of the significant covariates (at the 5% significance level) with right heart catheterization, we obtained a corresponding efficient estimate of 5.79% (SE 1.33%). These results lie close the previously obtained estimate of the population-averaged effect. This is not surprising because the shift of the propensity score distribution by weighting by $w(C)$ is relatively mild (Figure 5), and moreover, there is no evidence of treatment effect heterogeneity by the propensity score (P 0.74). These results

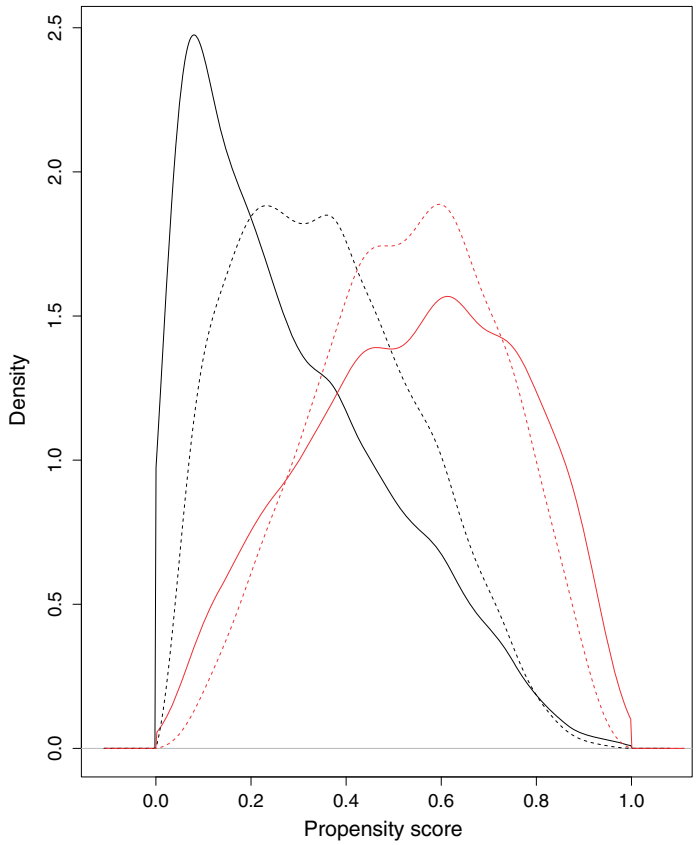

Figure 5. Data analysis. Solid: propensity score distribution for treated (red) and untreated (black) patients; dashed: (optimally) weighted propensity score distribution for treated (red) and untreated (black) patients (i.e. after weighting by $w(C)$).

express that the effect of right heart catheterization in a patient population about which treating physicians are less confident whether or not to apply right heart catheterization is to increase 30-day mortality risk with 5.79% (95% confidence interval 3.58% to 8.00%). For comparison, the inverse probability of treatment weighted estimator of the population-averaged effect (5.85% (SE 1.67%)) and corresponding double-robust estimator with calibration (5.34% (SE 1.56%)) (obtained using the R-package `iWeigReg` [24]) and without calibration (5.26% (SE 1.50%)) are of similar magnitude.

8. Discussion

We have shown that regression adjustment for the propensity score insulates standard tests of the causal null hypothesis in canonical generalised linear models against misspecification of the outcome model (provided that robust variance estimators are used). We have moreover shown that this robustness property extends to specific summaries, (7) and (9), of the effects obtained from these models, either on the difference or ratio scale. To the best of our knowledge, these results do not extend to weighted odds ratio estimators. We do not consider this to be disadvantageous in view of the more difficult interpretation of odds ratios and the fact that the weighted average exposure effect estimators (8) and (11) can still be based on logistic regression models to ensure that they respect the outcome range.

Our results on testing bear connections to recent findings by Rosenblum and van der Laan [25] that for a surprisingly large class of commonly used regression models that adjust for auxiliary covariates in randomised trials, standard hypothesis tests of no treatment effect are guaranteed to have correct type I error in large samples, even when the models are incorrectly specified (provided that robust variance estimators are utilised). Our results partially generalise these findings to observational data analysis based on regression adjustment for the propensity score.

Statistics in Medicine

The weighted average exposure effect estimators (8) and (11) based on regression adjustment for the propensity score inherit the simplicity as well as the properties of other propensity score approaches. The corresponding effect measures (7) and (9) differ, however, from those targeted by other propensity score approaches. Like other weighted effect measures (e.g. Mantel–Haenszel risk ratios), they reduce to the population-averaged effects $E(Y_1 - Y_0)$ and $E(Y_1)/E(Y_0)$, respectively, when the effects are homogenous across strata (in the sense that $E(Y_1 - Y_0|L) = \psi$ and $E(Y_1|L)/E(Y_0|L) = \psi$, respectively, for some constant ψ), but not otherwise. Without these homogeneity assumptions, the effect measures (7) and (9) may be considered less attractive from an interpretational point of view, because they express the treatment effect for a population different from the one under study. In particular, they express the treatment effect for a population obtained by eliminating individuals from the study population, who, on the basis of their covariate data, are nearly precluded from one of both treatment options. In a similar vein, Crump *et al*. [26] report effects for subpopulations of individuals who are sufficiently likely to receive both treatment options. With these authors [19, 26], we believe that the resulting effects may be of practical interest as they refer to a population of individuals for whom there is sufficient ambiguity about the optimal treatment option so as to render questions about treatment effects meaningful. In particular, the reweighted population may potentially be better resembling the typical clinical trial population, which excludes patients for whom, on the basis of their covariate data, one of both treatment options is considered unsuitable [18].

We believe that the weighted average exposure effects (7) and (9) are specifically of interest in settings with strong confounding. This is because they summarise the effects that prevail in those individuals in which the data carry substantial information about the exposure effect. In contrast, in such settings, population-averaged effects may sometimes attempt to answer a nearly irrelevant and overly ambitious causal question: how the population average outcome would look if everyone were exposed versus not exposed. That the data carry more information about the optimally weighted average exposure effects (7) than about the population-averaged effect was formally demonstrated by Crump *et al*. [19] in terms of a reduction of the semi-parametric efficiency bound; this reduction is especially pronounced in the presence of strong confounding.

We conclude by expressing our hope that we have developed a greater understanding of the properties of regression adjustment for propensity score and that the availability of optimally weighted average exposure effects that can be efficiently and robustly estimated sheds new light on the analysis of observational studies with strong confounding [27]. In future work, we will extend them to the analysis of time-varying treatments.

Appendix A

A.1. Properties of linear regression adjustment for the propensity score

It follows from results in [12] that the ordinary least squares estimator of θ_1 in model (3) is given by

$$\frac{\sum_{i=1}^{n}\left\{A_i - E^*(A_i|p(C_i))\right\}Y_i}{\sum_{i=1}^{n}\left\{A_i - E^*(A_i|p(C_i))\right\}A_i},$$

where $E^*(A_i|p(C_i))$ is the fitted value from a (population) least squares regression of A_i on $p(C_i)$. That is, $E^*(A_i|p(C_i)) = \alpha_0 + \alpha_1 p(C_i)$, with α_0 and α_1 the solution to the equations

$$\begin{aligned} 0 &= E\left\{A_i - \alpha_0 - \alpha_1 p(C_i)\right\} \\ 0 &= E\left[p(C_i)\left\{A_i - \alpha_0 - \alpha_1 p(C_i)\right\}\right]. \end{aligned}$$

It is readily verified that when the propensity score is correctly specified, the solution to this system of equations is given by $\alpha_0 = 0$ and $\alpha_1 = 1$, so that $E^*(A_i|p(C_i)) = p(C_i)$. It thus follows that the ordinary least squares estimator of θ_1 in model (3) is given by

$$\frac{\sum_{i=1}^{n}\left\{A_i - p(C_i)\right\}Y_i}{\sum_{i=1}^{n}\left\{A_i - p(C_i)\right\}A_i}.$$

If $E\left\{Y_1 - Y_0|p(C_i)\right\}$ is a constant, ψ, then it follows from the no unmeasured confounders Assumption (2) that $E\left\{Y|A=1,p(C_i)\right\} - E\left\{Y|A=0,p(C_i)\right\} = \psi$ and thus the conditional mean of the previous estimator, given the exposures and propensity scores for all individuals, equals

$$\frac{\sum_{i=1}^{n}\left\{A_i - p(C_i)\right\}E\left\{Y|A_i=0,p(C_i)\right\}}{\sum_{i=1}^{n}\left\{A_i - p(C_i)\right\}A_i} + \psi.$$

The numerator of the first term in the previous expression has mean zero, regardless of what is the true expectation $E\left\{Y|A=0,p(C_i)\right\}$, because of its multiplication by the mean zero residual $A_i - p(C_i)$. One can deduce from this that the ordinary least squares estimator of θ_1 in model (3) is a consistent estimator of ψ when $E\left\{Y_1 - Y_0|p(C_i)\right\}$ is constant and the propensity score is correctly specified, even when the outcome model (3) is misspecified.

A.2. Testing the causal null hypothesis

Suppose that the propensity score model is correctly specified so that $P(A=1|c) = p(c)$. Let furthermore $E(Y|A,C) = m(A,C;\theta^*)$, where $m(A,C;\theta)$ is the fitted value from a canonical generalised linear model with intercept and main effects in A and $p(C)$. In particular, $m(A,C;\theta)$ is thus a known function, smooth in θ and θ^* is an unknown finite-dimensional parameter; for example $m(A,C;\theta) = \text{expit}(\theta_0 + \theta_1 A + \theta_2' C)$. Let θ_1 be the exposure effect (i.e. the coefficient of A).

Suppose first that the propensity score is known. The score equations for θ in model (4) are

$$0 = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}1\\A_i\\p(C_i)\end{pmatrix}\left\{Y_i - m\left(A_i,C_i;\hat{\theta}\right)\right\}. \tag{A.1}$$

It can be deduced from this that in large samples, the maximum likelihood estimator $\hat{\theta}_1$ for θ_1 in model (4) satisfies

$$\sqrt{n}\left(\hat{\theta}_1 - \theta_1^*\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V^{-1}\left\{A_i - E^*(A_i|p(C_i))\right\}\left\{Y_i - m(A_i,C_i;\theta^*)\right\} + o_p(1),$$

where θ^* is the probability limit of $\hat{\theta}$, $o_p(1)$ is a variable that converges to zero in probability,

$$V = E\left[\left\{A_i - E^*(A_i|p(C_i))\right\}\nu(A_i,C_i)\right]$$

for $\nu(A_i,C_i) = \partial m\left(A_i,C_i;\theta^*\right)/\partial\theta_1$ and where $E^*(A_i|p(C_i))$ is the fitted value from a weighted (population) least squares regression of A_i on $p(C_i)$. That is, $E^*(A_i|p(C_i)) = \alpha_0 + \alpha_1 p(C_i)$, with α_0 and α_1 the solution to the equations

$$\begin{aligned}0 &= E\left[\nu(A_i,C_i)\left\{A_i - \alpha_0 - \alpha_1 p(C_i)\right\}\right]\\ 0 &= E\left[\nu(A_i,C_i)p(C_i)\left\{A_i - \alpha_0 - \alpha_1 p(C_i)\right\}\right].\end{aligned}$$

Under the causal null hypothesis,

$$\left\{A_i - E^*(A_i|p(C_i))\right\}\left\{Y_i - m(A_i,C_i;\theta^*)\right\}$$

has mean zero at $\theta_1^* = 0$. Indeed, in that case $\nu(A_i,C_i) = \nu(C_i)$ so that the solution to the previous system of equations is given by $\alpha_0 = 0$ and $\alpha_1 = 1$; hence, $E^*(A_i|p(C_i)) = p(C_i)$. It is now easily verified that the resulting estimating function

$$\left\{A_i - p(C_i)\right\}\left\{Y_i - m(A_i,C_i;\theta^*)\right\}$$

indeed has mean zero, regardless of the choice of $m(A_i,C_i;\theta^*)$ because $m(A_i,C_i;\theta^*)$ does not involve A_i when $\theta_1^* = 0$. It follows from the previous result that if the propensity score were known, then tests of

Statistics in Medicine

the causal null hypothesis are valid in the propensity score adjusted model (4), even when the outcome regression model is misspecified.

Consider now the more realistic case where the propensity score is unknown, but known to obey the regression model

$$P(A=1|C)=p(C;\gamma)=\text{expit}\left(\gamma_0+\gamma_1'C\right).$$

Suppose furthermore that this model is fitted using maximum likelihood so that an estimator $\hat{\gamma}$ of γ is obtained by solving

$$0=\frac{1}{n}\sum_{i=1}^{n}S_{\gamma,i}\equiv\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}1\\C_i\end{pmatrix}\left\{A_i-p(C_i;\hat{\gamma})\right\}.$$

Then, it follows by the usual Taylor series arguments that the asymptotic distribution of the maximum likelihood estimator $\hat{\theta}_1$ for the exposure effect $\theta_1^*=0$ in model (4), with $p(C)$ substituted by $p(C;\hat{\gamma})$, is given by

$$\begin{aligned}\sqrt{n}(\hat{\theta}_1-0)=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V^{-1}\left[\left\{A_i-p(C_i)\right\}\left\{Y_i-m\left(A_i,C_i;\theta^*\right)\right\}\right.\\&\left.-E\left[p(C)\left\{1-p(C)\right\}\begin{pmatrix}1\\C\end{pmatrix}'\left\{Y-m\left(A,C;\theta^*\right)\right\}\right]E\left(\partial S_\gamma/\partial\gamma\right)^{-1}S_{\gamma,i}\right]+o_p(1).\end{aligned}$$

When the outcome model is correctly specified, then

$$E\left[p(C)\left\{1-p(C)\right\}\begin{pmatrix}1\\C\end{pmatrix}'\left\{Y-m\left(A,C;\theta^*\right)\right\}\right]=0,$$

so that the uncertainty in the estimated propensity score can be ignored for inference about the exposure effect. When the outcome model is misspecified, then further insight can be developed upon defining $\varphi_i\equiv V^{-1}\left\{A_i-p(C_i)\right\}\left\{Y_i-m\left(A_i,C_i;\theta^*\right)\right\}$. With this notation, we now see that

$$\begin{aligned}\sqrt{n}(\hat{\theta}_1-0)&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_i-E\left(\partial\varphi_i/\partial\gamma\right)E\left(\partial S_{\gamma,i}/\partial\gamma\right)^{-1}S_{\gamma,i}+o_p(1)\\&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_i-E\left(\varphi_iS_{\gamma,i}'\right)E\left(S_{\gamma,i}S_{\gamma,i}'\right)^{-1}S_{\gamma,i}+o_p(1)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_i-\prod\left(\varphi_i|\Lambda_\gamma\right)+o_p(1),\end{aligned}$$

where $\prod\left(\varphi_i|\Lambda_\gamma\right)$ is the orthogonal projection of φ_i onto the linear span of all scores $S_{\gamma,i}$. Because the residual from such orthogonal projection varies less than the original variable, φ_i, we may conclude that a conservative test of the null hypothesis is obtained if one ignores the uncertainty in the estimated propensity score, provided that a robust (i.e. empirical) variance estimator is used. A more powerful test than the default test that ignores the uncertainty in the estimated propensity score can thus be based on the previous expansion of $\hat{\theta}_1$.

Consider now the extended model $P(Y=1|a,c)=\text{expit}\left\{\theta_0+\theta_1a+\theta_2p(c)+\theta_3'c\right\}$ with the propensity score assumed known. Then, by a similar reasoning as in the previous paragraph, we have that at the causal null hypothesis,

$$\sqrt{n}(\hat{\theta}_1-0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_0^{-1}\left\{A_i-p(C_i)\right\}\left\{Y_i-m\left(A_i,C_i;\theta^*\right)\right\}+o_p(1),\tag{A.2}$$

where $m(A_i,C_i;\theta^*)=m(0,C_i;\theta^*)=\text{expit}\left\{\theta_0^*+\theta_2^*p(C_i)+\theta_3^{*'}C_i\right\}$ because $\theta_1^*=0$; further,

$$V_0=E\left[\left\{A-p(C)\right\}m(0,C;\theta^*)\left\{1-m(0,C;\theta^*)\right\}A\right].$$

Because the term $\left\{A_i-p(C_i)\right\}$ in expression (A.2) remains unchanged upon adding covariates C to the model, the inclusion of additional covariates does not take away variation in the exposure. Because furthermore, the model-based variance of a binary outcome, that is, $m(0,C;\theta^*)\left\{1-m(0,C;\theta^*)\right\}$, is hardly

affected by the inclusion of additional covariates, we can deduce that at the causal null hypothesis, the inclusion of the baseline covariates in the propensity score adjusted model does not greatly affect the precision of the exposure effect estimator $\hat{\theta}_1$, unless it affects the correct specification of the outcome model. Indeed, under misspecification of the outcome model, it follows from (A.2) that the variance of $\hat{\theta}_1$ becomes

$$\frac{1}{nV_0^2} + \frac{E\left[p(C)\left\{1-p(C)\right\}\left\{E(Y|C) - m(0,C;\theta^*)\right\}^2\right]}{nV_0^2},$$

where the second term is the added variance because of model misspecification. Likewise, the variance of the directly standardised estimator $\hat{\psi}$, as given by (6), is expected to be only minimally affected by the additional inclusion of the covariates C in model (4) at the causal null hypothesis. This is because at the causal null hypothesis, the distribution of this estimator (with known propensity score) is given by

$$\sqrt{n}(\hat{\psi}-0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m\left(1,C_i;\theta^*\right) - m\left(0,C_i;\theta^*\right) + E\left[m(1,C_i;\theta^*)\left\{1-m(1,C_i;\theta^*)\right\}\right]\sqrt{n}\left(\hat{\theta}_1 - \theta_1^*\right) + o_p(1)$$

and thus only affected by the uncertainty in the estimation of θ_1.

A.3. Weighted effects on the difference scale

Suppose that the propensity score model is correctly specified so that $P(A = 1|c) = p(c)$. Denote $m(a,c;\theta) = \text{expit}\left\{\theta_0 + \theta_1 a + \theta_2 p(c)\right\}$. The score equations for θ in the propensity score adjusted model (4) are given by (A.1). We thus have that at the probability limit θ^* of $\hat{\theta}$, the following three restrictions must hold:

$$\begin{aligned} 0 &= E\{Y - m(A,C;\theta^*)\} = E\left[A\left\{E(Y|A=1,C) - m(1,C;\theta^*)\right\} + (1-A)\left\{E(Y|A=0,C) - m(0,C;\theta^*)\right\}\right] \\ &= E\left[p(C)\left\{E(Y|A=1,C) - m(1,C;\theta^*)\right\} + (1-p(C))\left\{E(Y|A=0,C) - m(0,C;\theta^*)\right\}\right] \end{aligned}$$

and

$$0 = E\left[A\{Y - m(A,C;\theta^*)\}\right] = E\left[A\{E(Y|A=1,C) - m(1,C;\theta^*)\}\right] = E\left[p(C)\left\{E(Y|A=1,C) - m(1,C;\theta^*)\right\}\right]$$

and

$$\begin{aligned} 0 &= E\left[p(C)\left\{Y - m(A,C;\theta^*)\right\}\right] \\ &= E\left[p(C)A\left\{E(Y|A=1,C) - m(1,C;\theta^*)\right\} + p(C)(1-A)\left\{E(Y|A=0,C) - m(0,C;\theta^*)\right\}\right] \\ &= E\left[p(C)^2\left\{E(Y|A=1,C) - m(1,C;\theta^*)\right\} + (1-p(C))p(C)\left\{E(Y|A=0,C) - m(0,C;\theta^*)\right\}\right]. \end{aligned}$$

In particular, we thus have that

$$\begin{aligned} E\left\{p(C)E(Y|A=1,C)\right\} &= E\left\{p(C)m(1,C;\theta^*)\right\} \\ E\left[\left\{1-p(C)\right\}E(Y|A=0,C)\right] &= E\left[\left\{1-p(C)\right\}m(0,C;\theta^*)\right] \\ E\left[p(C)^2\left\{E(Y|A=1,C) - E(Y|A=0,C)\right\}\right] &= E\left[p(C)^2\left\{m(1,C;\theta^*) - m(0,C;\theta^*)\right\}\right]. \end{aligned}$$

From these restrictions, it now follows that the large sample mean of (8) is

$$\begin{aligned} &E\left[p(C)\left\{1-p(C)\right\}\left\{m(1,C;\theta^*) - m(0,C;\theta^*)\right\}\right] / E\left[p(C)\left\{1-p(C)\right\}\right] \\ &= E\left[p(C)\left\{1-p(C)\right\}\left\{E(Y|A=1,C) - E(Y|A=0,C)\right\}\right] / E\left[p(C)\left\{1-p(C)\right\}\right] = E\left\{w(C)(Y_1 - Y_0)\right\}, \end{aligned}$$

even when the outcome model is misspecified.

A.4. Weighted effects on the ratio scale

The score equations for θ in the propensity score adjusted model (10) now involve the additional restriction

$$0 = \frac{1}{n}\sum_{i=1}^{n} A_i p(C_i)\left\{Y_i - m\left(A_i, C_i; \hat{\theta}\right)\right\},$$

which implies

$$\begin{aligned} 0 &= E\left[Ap(C)\left\{Y - m(A, C; \theta^*)\right\}\right] = E\left[p(C)A\left\{E(Y|A=1, C) - m(1, C; \theta^*)\right\}\right] \\ &= E\left[p(C)^2\left\{E(Y|A=1, C) - m(1, C; \theta^*)\right\}\right]. \end{aligned}$$

We thus have

$$\begin{aligned} E\left\{p(C)E(Y|A=1, C)\right\} &= E\left\{p(C)m(1, C; \theta^*)\right\} \\ E\left[\left\{1-p(C)\right\}E(Y|A=0, C)\right] &= E\left[\left\{1-p(C)\right\}m(0, C; \theta^*)\right] \\ E\left\{p(C)^2E(Y|A=1, C)\right\} &= E\left\{p(C)^2 m(1, C; \theta^*)\right\} \\ E\left\{p(C)\left\{1-p(C)\right\}E(Y|A=0, C)\right\} &= E\left\{p(C)\left\{1-p(C)\right\}m(0, C; \theta^*)\right\}. \end{aligned}$$

From these restrictions, it now follows that the large sample mean of

$$\begin{aligned} &E\left[p(C)\left\{1-p(C)\right\}m(1, C; \theta^*)\right] / E\left[p(C)\left\{1-p(C)\right\}\right] \\ &= E\left[p(C)\left\{1-p(C)\right\}E(Y|A=1, C)\right] / E\left[p(C)\left\{1-p(C)\right\}\right] = E\left\{w(C)Y_1\right\}, \end{aligned}$$

even when the outcome model is misspecified and likewise.

$$E\left[p(C)\left\{1-p(C)\right\}m(0, C; \theta^*)\right] / E\left[p(C)\left\{1-p(C)\right\}\right] = E\left\{w(C)Y_0\right\}.$$

Acknowledgements

The first author acknowledges support from IAP research network grant no. P07/06 from the Belgian government (Belgian Science Policy). The second author acknowledges support from the U.K. Medical Research Council (Career Development Award in Biostatistics, G1002283).

References

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
2. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998; **17**:2265–2281.
3. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
4. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997; **127**:757–763.
5. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1984; **39**:33–38.
6. Robins JM. Marginal structural models. *Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, 1997, 1-10.
7. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**:550–560.
8. Tan Z. Understanding, OR, PS, and DR. *Statistical Science* 2008; **22**:560–568.
9. Rubin DB. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 1990; **25**:279–292.
10. Hernán MA. A definition of causal effect for epidemiological research. *Journal of Epidemiology and Community Health* 2004; **58**:265–271.
11. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology* 2009; **20**:880–883.
12. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 1992; **48**:479–495.

13. Greenland S, Robins J, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**:29–46.
14. Stampf S, Graf E, Schmoor C, Schumacher M. Estimators and confidence intervals for the marginal odds ratio using logistic regression and propensity score stratification. *Statistics in Medicine* 2010; **29**:760–769.
15. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology* 2011; **173**:731–738.
16. Vansteelandt S, Keiding N. Invited commentary: G-computation – lost in translation *American Journal of Epidemiology* 2011; **173**:739–742.
17. Robinson L, Jewell N. Some surprising results about covariate adjustment in logistic-regression models. *International Statistical Review* 1991; **59**:227–240.
18. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* 2006; **163**:262–270.
19. Crump RV, Hotz J, Imbens I, Mitnik O. Moving the goalposts: addressing limited overlap in the estimation of average treatment effects by changing the estimand, Technical Report No. 330, NBER Technical Working Paper 2006.
20. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution – a simulation study. *American Journal of Epidemiology* 2010; **172**:843–854.
21. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 2012; **21**:31–54.
22. Connors AF, Speroff T, Dawson NV, Thomas C, Harrell FE, Wagner D, Desbiens N, Goldman L, Wu AW, Califf RM, Fulkerson WJ, Vidaillet H, Broste S, Bellamy P, Lynn J, Knaus WA. The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association* 1996; **276**:889–897.
23. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcome Research Methods* 2001; **2**:259–278.
24. Tan Z. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* 2006; **101**:1619–1637.
25. Rosenblum M, van der Laan MJ. 2009. Using regression models to analyze randomized trials: asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics* 2009; **65**:937-945.
26. Crump RV, Hotz J, Imbens I, Mitnik O. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 2009; **96**:187–199.
27. Moore KL, Neugebauer R, van der Laan MJ, Tager IB. Causal inference in epidemiological studies with strong confounding. *Statistics in Medicine* 2012; **31**:1380–1404.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.