STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:2937–2960 (DOI: 10.1002/sim.1903)

Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study

Jared K. Lunceford[1,*,†] and Marie Davidian[2]

[1]*Merck Research Laboratories, RY34-A316, P.O. Box 2000, Rahway, NJ 07065-0900, U.S.A.*
[2]*Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695, U.S.A.*

SUMMARY

Estimation of treatment effects with causal interpretation from observational data is complicated because exposure to treatment may be confounded with subject characteristics. The propensity score, the probability of treatment exposure conditional on covariates, is the basis for two approaches to adjusting for confounding: methods based on stratification of observations by quantiles of estimated propensity scores and methods based on weighting observations by the inverse of estimated propensity scores. We review popular versions of these approaches and related methods offering improved precision, describe theoretical properties and highlight their implications for practice, and present extensive comparisons of performance that provide guidance for practical use. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: covariate balance; double robustness; inverse-probability-of-treatment-weighted-estimator; observational data

1. INTRODUCTION

Observational data are often the basis for epidemiological and other investigations seeking to make inference on the effect of treatment exposure on a response. Randomized studies aim to balance distributions of subject characteristics across groups, so that groups are similar except for the treatments. However, with observational data, treatment exposure may be associated with covariates that are also associated with potential response, and groups may be seriously imbalanced in these factors. Consequently, unbiased treatment comparisons from observational data require methods that adjust for such confounding of exposure to treatment with subject characteristics, and inferences with a causal interpretation cannot be made without appropriate adjustment.

*Correspondence to: Jared K. Lunceford, Merck Research Laboratories, RY34-A316, P.O. Box 2000, Rahway, NJ 07065-0900, U.S.A.
†E-mail: jared_lunceford@merck.com

Contract/grant sponsor: NIH; contract/grant numbers: R01-CA085848 and R37-AI031789

Received September 2003
Accepted April 2004

For comparing two treatments, 'treated' and 'control', say, the propensity score is the probability of exposure to treatment conditional on observed covariates [1]. Properties of the propensity score that facilitate causal inferences are given by Rosenbaum and Rubin [1] (see also References [2, 3]), and applications of methods using adjustments based on propensity scores are increasingly widespread, e.g. References [4–6]. A popular method for estimating the (causal) difference of two treatment means is that of Rosenbaum and Rubin [7], where individuals are stratified based on estimated propensity scores and the difference estimated as the average of within-stratum effects. An alternative approach is to adjust for confounding by using estimated propensity scores to construct weights for individual observations [8, 9].

In this paper, we review approaches using stratification and weighting based on propensity scores for making causal inferences from observational data and contrast their performance. A main objective is to provide a mostly self-contained introduction to these methods and their underpinnings, a description of their properties that highlights insights with implications for practice, and a demonstration of relative performance that suggests guidelines for application. In Section 2, we discuss the framework of counterfactuals or potential outcomes [10], which formalizes the notion of 'causal effect,' and assumptions required to justify adjustments for confounding. We describe popular propensity-score-based approaches and describe some additional methods that may be less familiar to practitioners that may improve upon these. Section 3 presents theoretical properties of the estimators, and Section 4 reports on extensive comparative simulations.

2. ESTIMATORS BASED ON THE PROPENSITY SCORE

2.1. *Counterfactual framework*

Let Z be an indicator of observed treatment exposure ($Z = 1$ if treated, $Z = 0$ if control) and $\mathbf{X}$ be a vector of covariates measured prior to receipt of treatment (baseline) or, if measured post-treatment, not affected by either treatment. Each individual is assumed to have an associated random vector (Y_0, Y_1), where Y_0 and Y_1 are the values of the response that would be seen if, possibly contrary to the fact of what actually happened, s/he were to receive control or treatment, respectively. Consequently, Y_0 and Y_1 are referred to as counterfactuals (or potential outcomes) and may be viewed as inherent characteristics of the individual. The response Y *actually observed* is assumed to be that would be seen under the exposure actually received, formalized as

$$Y = Y_1 Z + (1 - Z) Y_0 \tag{1}$$

Thus, $(Y, Z, \mathbf{X})$ are observed on each individual. It is important to distinguish between the *observed* response Y and the *counterfactual* responses Y_0 and Y_1. The latter are hypothetical and may never be observed simultaneously; however, they are a convenient construct allowing precise statement of questions of interest, as we now describe.

The distributions of Y_0 and Y_1 may be thought of as representing the hypothetical distributions of response for the population of individuals were all individuals to receive control or be treated, respectively, so the means of these distributions correspond to the mean response if all individuals were to receive each treatment. Hence, a difference in these means would

be attributable to, or *caused by*, the treatments. Formally, then,

$$\Delta = \mu_1 - \mu_0 = E(Y_1) - E(Y_0)$$

is referred to as the average causal effect (of the treated state relative to control). Estimation of Δ is thus of central interest in comparing treatments.

This framework makes it possible to formalize the difficulty in estimating Δ, and thus making causal statements, from observational data. The counterfactuals are never both observed for any subject; thus, whether estimation of Δ is possible relies on whether $E(Y_0)$ and $E(Y_1)$ may be identified from the observed data $(Y, Z, \mathbf{X})$. The sample average response in the treated group estimates $E(Y \mid Z = 1)$, the mean of observed responses among subjects observed to be treated, which from (1) is equal to $E(Y_1 \mid Z = 1)$ but is different from $E(Y_1)$, the mean if the entire population were treated, and similarly for control. In a randomized trial, as Z is determined for each participant at random, it is unrelated to how s/he might *potentially respond*, and thus $(Y_0, Y_1) \perp\!\!\!\perp Z$, where $\perp\!\!\!\perp$ denotes statistical independence. Here, using (1), we thus have $E(Y \mid Z = 1) = E(Y_1 \mid Z = 1) = E(Y_1)$, and similarly $E(Y \mid Z = 0) = E(Y_0)$, verifying that the sample average difference is an unbiased estimator for Δ with a causal interpretation, as widely accepted. However, in an observational study, because treatment exposure Z is not controlled, Z may not be independent of (Y_0, Y_1); indeed, the same characteristics that lead an individual to be exposed to a treatment may also be associated, or 'confounded,' with his/her potential response. In this case, $E(Y \mid Z = 1) = E(Y_1 \mid Z = 1) \neq E(Y_1)$ and $E(Y \mid Z = 0) = E(Y_0 \mid Z = 0) \neq E(Y_0)$, so that the difference of observed sample averages is not an unbiased estimator for Δ. It is important to distinguish between the conditions $(Y_0, Y_1) \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$. The former involves *potential responses*, which are indeed independent of treatment assignment under randomization, while the latter involves the *observed response* and is unlikely to be true under any circumstances unless treatment has no effect.

In an observational study, although $(Y_0, Y_1) \perp\!\!\!\perp Z$ is unlikely to hold, it may be possible to identify subject characteristics related to both potential response and treatment exposure, referred to as 'confounders.' If we believe that $\mathbf{X}$ contains all such confounders, then, for individuals sharing a particular value of $\mathbf{X}$, there would be no association between the exposure states and the values of potential responses; i.e. treatment exposure among individuals with a particular $\mathbf{X}$ is essentially at random. Formally, Y_0, Y_1 are independent of treatment exposure conditional on $\mathbf{X}$, written

$$(Y_0, Y_1) \perp\!\!\!\perp Z \mid \mathbf{X} \tag{2}$$

Rosenbaum and Rubin [1] refer to (2) as the assumption of strongly ignorable treatment assignment; (2) has also been called the assumption of no unmeasured confounders [9]. One must appreciate that (2) is an *assumption*; willingness to assume (2) requires the analyst to have confidence that $\mathbf{X}$ contains all characteristics related to both treatment and response and that there are no additional, unmeasured such confounders.

The benefit of (2) is that $E(Y_0)$ and $E(Y_1)$ may be identified from $(Y, Z, \mathbf{X})$. The regression relationship $E(Y \mid Z, \mathbf{X})$ depends only on the observed data, so is identifiable. Then the average for $Z = 1$ over all $\mathbf{X}$ satisfies $E\{E(Y \mid Z = 1, \mathbf{X})\} = E\{E(Y_1 \mid Z = 1, \mathbf{X})\} = E\{E(Y_1 \mid \mathbf{X})\} = E(Y_1)$, where the first equality is from (1), the second follows from (2), and the outer expectation is with respect to the distribution of $\mathbf{X}$; similarly, $E\{E(Y \mid Z = 0, \mathbf{X})\} = E(Y_0)$. Thus, it should

be possible to make inferences on Δ if (2) may be assumed to hold. Methods using the propensity score are one way to achieve this.

2.2. The propensity score

The propensity score $e(\mathbf{X}) = P(Z = 1 \mid \mathbf{X})$, $0 < e(\mathbf{X}) < 1$, is the probability of treatment given the observed covariates. Rosenbaum and Rubin [1] showed that $\mathbf{X} \perp\!\!\!\perp Z \mid e(\mathbf{X})$, so individuals from either treatment group with the same propensity score are 'balanced' in that the distribution of $\mathbf{X}$ is the same regardless of exposure status. Rosenbaum and Rubin show that if (2) holds, in addition $(Y_0, Y_1) \perp\!\!\!\perp Z \mid e(\mathbf{X})$, so that treatment exposure is unrelated to the counterfactuals for individuals sharing the same propensity score. We now review ways these developments may be exploited to derive estimators for Δ from observed data $(Y_i, Z_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, an i.i.d. sample containing both treated and control subjects.

In practice, the propensity score is unlikely to be known, so it is routine to estimate it from the observed data $(Z_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, by assuming that $e(\mathbf{X})$ follows a parametric model, e.g. a logistic regression model $e(\mathbf{X}, \boldsymbol{\beta}) = \{1 + \exp(-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})\}^{-1}$, $\boldsymbol{\beta}(p \times 1)$. Interaction and higher-order terms may also be included. Here, $\boldsymbol{\beta}$ may be estimated by the maximum likelihood (ML) estimator $\tilde{\boldsymbol{\beta}}$ solving

$$\sum_{i=1}^{n} \psi_\beta(Z_i, \mathbf{X}_i, \boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{Z_i - e(\mathbf{X}_i, \boldsymbol{\beta})}{e(\mathbf{X}_i, \boldsymbol{\beta})\{1 - e(\mathbf{X}_i, \boldsymbol{\beta})\}} \, \partial/\partial\boldsymbol{\beta}\{e(\mathbf{X}_i, \boldsymbol{\beta})\} = \mathbf{0} \tag{3}$$

We assume that the analyst is proficient at modelling $e(\mathbf{X}, \boldsymbol{\beta})$, so that it is correctly specified, and write $e = e(\mathbf{X}, \boldsymbol{\beta})$ and $e_\beta = \partial/\partial\boldsymbol{\beta}\{e(\mathbf{X}, \boldsymbol{\beta})\}$, with subscript i when evaluated at $\mathbf{X}_i$.

2.3. Estimation of Δ based on stratification

The popular approach using stratification on estimated propensity scores to estimate Δ involves the following steps: (i) Estimate $\boldsymbol{\beta}$ as in (3) and calculate estimated propensity scores $\hat{e}_i = e(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$ for all i; (ii) form K strata according to the sample quantiles of the $\hat{e}_i$, where the jth sample quantile $\hat{q}_j$, $j = 1, \ldots, K$, is such that the proportion of $\hat{e}_i \leqslant \hat{q}_j$ is roughly j/K, $\hat{q}_0 = 0$, and $\hat{q}_K = 1$; (iii) within each stratum, calculate the difference of sample means of the Y_i for each treatment; and (iv) estimate Δ by a weighted sum of the differences of sample means across strata, where weighting is by the proportion of observations falling in each stratum. Defining $\hat{Q}_j = (\hat{q}_{j-1}, \hat{q}_j]$; $n_j = \sum_{i=1}^{n} I(\hat{e}_i \in \hat{Q}_j)$, the number of individuals in stratum j; and $n_{1j} = \sum_{i=1}^{n} Z_i I(\hat{e}_i \in \hat{Q}_j)$ is the number of these who are treated, the estimator using a weighted sum is

$$\hat{\Delta}_S = \sum_{j=1}^{K} \left(\frac{n_j}{n}\right) \left\{ n_{1j}^{-1} \sum_{i=1}^{n} Z_i Y_i I(\hat{e}_i \in \hat{Q}_j) - (n_j - n_{1j})^{-1} \sum_{i=1}^{n} (1 - Z_i) Y_i I(\hat{e}_i \in \hat{Q}_j) \right\} \tag{4}$$

As the weights $n_j/n \approx K^{-1}$, they may be replaced by K^{-1} to yield an average across strata.

The rationale follows from the property $(Y_0, Y_1) \perp\!\!\!\perp Z \mid e(\mathbf{X})$ when (2) holds. Because treatment exposure is essentially at random for individuals with the same propensity value, we expect mean comparisons within this group to be unbiased. Identifying individuals sharing exactly the same propensity value may be infeasible in practice, so stratification attempts to achieve groups where this at least holds approximately. Consequently, $\hat{\Delta}_S$ may be a biased

estimator of Δ, as some residual confounding within strata may remain. Rosenbaum and Rubin [1, 7] advocate the use of quantiles ($K = 5$), a choice made in most published applications. Intuitively, these results require that the propensity model be correctly specified. Thus, it is often recommended [5, 7] that, following (ii), the analyst examine the degree of balance for each element of $\mathbf{X}$ within each stratum using standard statistical tests. Evidence that balance has not been achieved may reflect an incorrect model and the need for refinement, followed by a return to (i).

To reduce residual within-stratum confounding, a variation on (4) is often advocated [2, 11]. Here, steps (iii) and (iv) are modified as follows: (iii) within each stratum $j = 1, \ldots, K$, fit a regression model of the form $m^{(j)}(Z, \mathbf{X}, \boldsymbol{\alpha}^{(j)})$ representing the postulated regression relationship $E(Y \mid Z, \mathbf{X})$ within stratum j and, based on the resulting estimate $\hat{\boldsymbol{\alpha}}^{(j)}$, estimate treatment effect in stratum j by averaging over $\mathbf{X}_i$ in j as

$$\hat{\Delta}^{(j)} = n_j^{-1} \sum_{i=1}^{n} I(\hat{e}_i \in \hat{Q}_j)\{m^{(j)}(1, \mathbf{X}_i, \hat{\boldsymbol{\alpha}}^{(j)}) - m^{(j)}(0, \mathbf{X}_i, \hat{\boldsymbol{\alpha}}^{(j)})\} \tag{5}$$

and (iv) estimate Δ by the average or weighted sum of the $\hat{\Delta}^{(j)}$, e.g. using the average

$$\hat{\Delta}_{SR} = K^{-1} \sum_{j=1}^{K} \hat{\Delta}^{(j)} \tag{6}$$

Ordinarily, the $m^{(j)}$ are taken to be the same function of Z and $\mathbf{X}$ for all j. E.g. for a linear model, $m^{(j)}(Z, \mathbf{X}, \boldsymbol{\alpha}^{(j)}) = \alpha_0^{(j)} + \alpha_Z^{(j)} Z + \mathbf{X}^{\mathrm{T}} \boldsymbol{\alpha}_X^{(j)}$; here, $\hat{\Delta}^{(j)} = \hat{\alpha}_Z^{(j)}$ for each j.

Within-stratum regression modelling is intended to eliminate any remaining imbalances within strata. In Section 3.2, we demonstrate that while $\hat{\Delta}_S$ does not yield a consistent estimator for Δ in general, $\hat{\Delta}_{SR}$ is consistent as long as the models $m^{(j)}$ all coincide with the true, overall regression relationship $E(Y \mid Z, \mathbf{X})$, but may be inconsistent otherwise.

2.4. Estimation of Δ based on weighting

Rather than seeking unbiased estimation within strata, weighting methods attempt to obtain an unbiased estimator for Δ in a way akin to that proposed by Horvitz and Thompson [12]. Under (1), as $Z(1 - Z) = 0$, $E\{ZY/e(\mathbf{X})\} = E\{ZY_1/e(\mathbf{X})\}$, so that, assuming (2),

$$E\left\{\frac{ZY}{e(\mathbf{X})}\right\} = E\left[E\left\{\frac{I(Z=1)Y_1}{e(\mathbf{X})}\,\middle|\, Y_1, \mathbf{X}\right\}\right] = E\left\{\frac{Y_1}{e(\mathbf{X})} E\{I(Z=1) \mid Y_1, \mathbf{X}\}\right\} = E(Y_1)$$

where (2) implies $E\{I(Z=1) \mid Y_1, \mathbf{X}\} = e(\mathbf{X})$, allowing the last equality; and we have used $Z = I(Z=1)$. Similarly, $E[(1 - Z)Y/\{1 - e(\mathbf{X})\}] = E(Y_0)$. This suggests immediately the estimator for Δ proposed by Rosenbaum [3] and others

$$\hat{\Delta}_{IPW1} = n^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i}{\hat{e}_i} - n^{-1} \sum_{i=1}^{n} \frac{(1 - Z_i) Y_i}{1 - \hat{e}_i} = \hat{\mu}_{1,IPW1} - \hat{\mu}_{0,IPW1} \tag{7}$$

$E\{Z/e(\mathbf{X})\} = E\{E(Z \mid \mathbf{X})/e(\mathbf{X})\} = 1$ and $E[(1-Z)/\{1-e(\mathbf{X})\}] = 1$ suggest

$$\hat{\Delta}_{IPW2} = \left(\sum_{i=1}^{n} \frac{Z_i}{\hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i}{\hat{e}_i} - \left(\sum_{i=1}^{n} \frac{1-Z_i}{1-\hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{(1-Z_i)Y_i}{1-\hat{e}_i} = \hat{\mu}_{1,IPW2} - \hat{\mu}_{0,IPW2} \tag{8}$$

The estimator for a single mean in (8) is known as a ratio estimator in the sampling literature.

As (7) and (8) involve weighting the observations in each group by the inverse of the probability of being in that group, '*IPW*' denotes 'inverse probability weighting,' and $\hat{\Delta}_{IPW1}$ and $\hat{\Delta}_{IPW2}$ are popular approaches based on such weighting. However, they are special cases of a broader class of estimators that may be deduced by viewing the situation as a 'missing data' problem discussed in a landmark paper by Robins, Rotnitzky, and Zhao [13]. To appreciate this, consider μ_1. Identifying $(Y_1, Z, \mathbf{X})$ as the 'full data,' Y_1 is only observed for individuals with $Z = 1$ (and is 'missing' for those with $Z = 0$), so that the probability of a 'complete case' is $P(Z = 1 \mid \mathbf{X})$ if treatment is related to $\mathbf{X}$. Inverse weighting in the first terms of $\hat{\Delta}_{IPW1}$ and $\hat{\Delta}_{IPW2}$ allows each 'complete case' i to count for him/herself and $(\hat{e}_i^{-1} - 1)$ other 'missing' subjects with like characteristics $\mathbf{X}_i$ in estimating μ_1. From this 'missing data' perspective, the Robins *et al.* theory may be used to describe the class of all consistent, semiparametric estimators for μ_1 and μ_0 and hence Δ; i.e. estimators that do not require the distribution of $(Y_1, Y_0, \mathbf{X})$ to be specified. The theory shows that all such estimators for Δ involve 'inverse weighting' of 'complete cases' and are consistent if the complete-case probability (i.e. the propensity score) is correctly modeled, so should be approximately unbiased in finite samples. The class includes simple estimators such as $\hat{\Delta}_{IPW1}$ and $\hat{\Delta}_{IPW2}$ [for μ_0, the complete-case probability is $P(Z = 0 \mid \mathbf{X}) = 1 - P(Z = 1 \mid \mathbf{X})$], but others are possible. We describe two alternative estimators here.

The theory of Robins *et al.* [13] identifies the estimator within the class having the smallest (large-sample) variance, the (locally) semiparametric efficient estimator

$$\hat{\Delta}_{DR} = n^{-1} \sum_{i=1}^{n} \frac{Z_i Y_i - (Z_i - \hat{e}_i) m_1(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_1)}{\hat{e}_i} - n^{-1} \sum_{i=1}^{n} \frac{(1-Z_i)Y_i + (Z_i - \hat{e}_i) m_0(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_0)}{1-\hat{e}_i} \tag{9}$$

Here $m_z(\mathbf{X}, \boldsymbol{\alpha}_z) = E(Y \mid Z = z, \mathbf{X})$ is the regression of the response on $\mathbf{X}$ in group z, $z = 0, 1$, depending on parameters $\boldsymbol{\alpha}_z$, and $\hat{\boldsymbol{\alpha}}_z$ is an estimator for $\boldsymbol{\alpha}_z$ based on the data from subjects with $Z = z$. Each term in $\hat{\Delta}_{DR}$ has the form of those in $\hat{\Delta}_{IPW1}$ and $\hat{\Delta}_{IPW2}$ but 'augmented' (e.g. Reference [14]) by an expression involving the regression; it is this 'augmentation' that serves to increase efficiency. Unlike $\hat{\Delta}_S$, $\hat{\Delta}_{IPW1}$, and $\hat{\Delta}_{IPW2}$, $\hat{\Delta}_{DR}$ requires specification of this regression model; however, because $\hat{\Delta}_{DR}$ is the efficient estimator in the class, in large samples, it has smaller variance than $\hat{\Delta}_{IPW1}$ or $\hat{\Delta}_{IPW2}$, often dramatically so. Moreover, Scharfstein *et al.* [15, Section 3.2.3] note that $\hat{\Delta}_{DR}$ has a so-called 'double-robustness' property that the estimator remains consistent if either (i) the propensity score model is correctly specified but the two regression models m_0 and m_1 are not or (ii) the two regression models are correctly specified but the propensity score model is not, although under these conditions it need no longer be most efficient. Neither $\hat{\Delta}_{IPW1}$ nor $\hat{\Delta}_{IPW2}$ need be consistent if e is incorrectly specified, as the motivating arguments earlier in this section would no longer be valid.

It is also possible to derive other estimators in the Robins *et al.* class that do not incorporate regression modeling by attempting to improve directly upon estimation of μ_1 and μ_0.

With $\boldsymbol{\beta}$ known, the estimators for μ_1 and μ_0 in $\hat{\Delta}_{IPW1}$ and $\hat{\Delta}_{IPW2}$ solve

$$\sum_{i=1}^{n}\left\{\frac{Z_i(Y_i-\mu_1)}{e_i}+\eta_1\left(\frac{Z_i-e_i}{e_i}\right)\right\}=0 \quad \text{and} \quad \sum_{i=1}^{n}\left\{\frac{(1-Z_i)(Y_i-\mu_0)}{1-e_i}-\eta_0\left(\frac{Z_i-e_i}{1-e_i}\right)\right\}=0 \tag{10}$$

respectively, where $(\eta_0,\eta_1)=(\mu_0,\mu_1)$ yields $\hat{\Delta}_{IPW1}$ and $(\eta_0,\eta_1)=(0,0)$ gives $\hat{\Delta}_{IPW2}$. This suggests improving upon $\hat{\Delta}_{IPW1}$ and $\hat{\Delta}_{IPW2}$ by identifying constants η_0,η_1 that minimize the large-sample variance of solutions to the equations in (10), given by $\eta_1=-E\{Z(Y-\mu_1)/e^2\}/E\{(Z-e)^2/e^2\}$ and $\eta_0=-E\{(1-Z)(Y-\mu_0)/(1-e)^2\}/E\{(Z-e)^2/(1-e)^2\}$, which motivates estimating these constants by solving

$$\sum_{i=1}^{n}\left\{\frac{(Z_i(Y_i-\mu_1)}{e_i^2}+\eta_1\left(\frac{Z_i-e_i}{e_i}\right)^2\right\}=0 \text{ and } \sum_{i=1}^{n}\left\{\frac{(1-Z_i)(Y_i-\mu_0)}{(1-e_i)^2}+\eta_0\left(\frac{Z_i-e_i}{1-e_i}\right)^2\right\}=0 \tag{11}$$

In practice, one would estimate $\boldsymbol{\beta}$, solving (10) and (11) jointly with (3), yielding

$$\begin{aligned}\hat{\Delta}_{IPW3}&=\left\{\sum_{i=1}^{n}\frac{Z_i}{\hat{e}_i}\left(1-\frac{C_1}{\hat{e}_i}\right)\right\}^{-1}\sum_{i=1}^{n}\frac{Z_iY_i}{\hat{e}_i}\left(1-\frac{C_1}{\hat{e}_i}\right)\\&\quad-\left\{\sum_{i=1}^{n}\frac{1-Z_i}{1-\hat{e}_i}\left(1-\frac{C_0}{1-\hat{e}_i}\right)\right\}^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}_i}\left(1-\frac{C_0}{1-\hat{e}_i}\right)\\&=\hat{\mu}_{1,IPW3}-\hat{\mu}_{0,IPW3}\end{aligned} \tag{12}$$

$$C_1=\sum_{i=1}^{n}\{(Z_i-\hat{e}_i)/\hat{e}_i\}\Big/\sum_{i=1}^{n}\{(Z_i-\hat{e}_i)/\hat{e}_i\}^2$$

$$C_0=-\sum_{i=1}^{n}\{(Z_i-\hat{e}_i)/(1-\hat{e}_i)\}\Big/\sum_{i=1}^{n}\{(Z_i-\hat{e}_i)/(1-\hat{e}_i)\}^2$$

Unlike (7) and (8), in the first term of (12), each weight $\hat{e}_i^{-1}$ is proportionately scaled by a measure of how the sample, weighted exposure indicators $Z_i/\hat{e}_i$ deviate from their expectation (if $\boldsymbol{\beta}$ were known) of 1, and similarly for the second term. In large samples, C_0, C_1 should be close to 0, but for smaller n, this scaling proportionately reduces or increases each 'complete-case' weight. For $\hat{\Delta}_{IPW1}$ and $\hat{\Delta}_{IPW2}$, inverse weighting an observation by a very small complete-case probability can result in numerical instability, particularly when n is not large. Thus, the scaling has the effect in practice of offering stability in the case where some complete-case probabilities may be small or are highly variable. Interestingly, the 'augmentation' incorporated in $\hat{\Delta}_{DR}$ tends to lessen such instability problems in practice.

As we demonstrate in Section 4, estimators like $\hat{\Delta}_{IPW3}$ that do not incorporate regression models, although improving in precision over $\hat{\Delta}_{IPW1}$ and $\hat{\Delta}_{IPW2}$, cannot achieve the efficiency

gains possible through 'augmentation' involving regression as in $\hat{\Delta}_{DR}$. Hirano and Imbens [16] report on a practical application of weighted methods and advocate incorporation of regression models as in (9) for this reason.

2.5. Summary

It is important to recognize that incorporation of regression modelling in $\hat{\Delta}_{DR}$ and $\hat{\Delta}_{SR}$ is different from a popular alternative to all estimators previously discussed, that of estimation of Δ *directly* from a regression model. For example, for a linear model $E(Y \mid Z, \mathbf{X}) = \alpha_0 + \alpha_Z Z + \mathbf{X}^{\mathrm{T}}\alpha_X$, under (2), it is straightforward to verify that $\Delta = E(Y_1) - E(Y_0) = E\{E(Y \mid Z = 1, \mathbf{X})\} - E\{E(Y \mid Z = 0, \mathbf{X})\} = \alpha_Z$. For models nonlinear in $\mathbf{X}$ such as the logistic, this difference may not have a closed form, as each term involves integration over the distribution of $\mathbf{X}$. In either case, the direct modelling approach has serious drawbacks; Rubin [17] offers an excellent discussion. When $\dim(\mathbf{X})$ is large, ensuring that the regression model is correct, and hence that a consistent estimator for Δ will be obtained, is difficult. In addition, if the distributions of some confounders do not overlap substantially in the treated and control groups, the regression relationship is determined primarily by treated subjects in one region of the $\mathbf{X}$ space and by control subjects in another, so that estimates of causal effects using direct modelling are essentially based on extrapolation. In contrast, the regression modelling used by $\hat{\Delta}_{SR}$ largely circumvents this, as $\mathbf{X}$ and Z should be approximately independent within-strata. Moreover, by 'double robustness,' even if the regression models in $\hat{\Delta}_{DR}$ are incorrect, this estimator, which incorporates regression models only as a way to gain efficiency over simpler weighted estimators, will still be consistent.

When the true regression is linear and $\mathrm{var}(Y \mid Z, \mathbf{X})$ is constant, direct modelling may be implemented by ordinary least squares (*OLS*), which is ML estimation if $Y \mid Z, \mathbf{X}$ has a normal distribution. If, in fact, these conditions hold, and the chosen model for $E(Y \mid Z, \mathbf{X})$ is correctly specified by the analyst, then standard large sample theory implies that the resulting estimator for Δ will be consistent and the most efficient. One would thus expect the direct regression approach to outperform those based on propensity scores; however, such gains would be at the risk of the disadvantages noted above. In Section 4, we investigate these issues empirically. The same considerations apply to ML estimation for any regression model, e.g., logisitic regression for binary response.

As noted, $\hat{\Delta}_{IPW1}$, $\hat{\Delta}_{IPW2}$, $\hat{\Delta}_{IPW3}$, and $\hat{\Delta}_{DR}$ are all members of the class of consistent, semiparametric estimators of Robins *et al.* [13]. However, as shown in Section 3.2, for fixed K, $\hat{\Delta}_S$ is not consistent and evidently neither $\hat{\Delta}_S$ nor $\hat{\Delta}_{SR}$ makes use of inverse weighting, so these estimators are not members of this class. Thus, although insights into additional properties of $\hat{\Delta}_{IPW1}$, $\hat{\Delta}_{IPW2}$, $\hat{\Delta}_{IPW3}$, and $\hat{\Delta}_{DR}$ follow easily from the Robins *et al.* theory, as shown next in Sections 3.1 and 3.3, those for $\hat{\Delta}_S$ and $\hat{\Delta}_{SR}$ must be deduced separately.

3. THEORETICAL PROPERTIES

In this section we summarize properties of the estimators and highlight the practical insights that can be deduced from these. The large-sample properties for weighted estimators follow from the general framework of Reference [13] and may also be obtained directly from the

standard theory of M-estimation, as we describe in Section 3.1. The properties for stratification estimators to our knowledge have not been elucidated and are sketched in Section 3.2.

3.1. Weighted estimators

Properties of $\hat{\Delta}_{IPW1}$, $\hat{\Delta}_{IPW2}$, $\hat{\Delta}_{IPW3}$, and $\hat{\Delta}_{DR}$ when e is correctly specified may be deduced by viewing them as solutions to a set of estimating equations. To obtain $\hat{\Delta}_{IPW1}$ and $\hat{\Delta}_{IPW2}$, one solves jointly in $(\Delta, \boldsymbol{\beta})$ (3) and an equation of the form $\sum_{i=1}^{n} \psi_{\Delta}(Y_i, Z_i, \mathbf{X}_i, \Delta, \boldsymbol{\beta}) = 0$ that follows from (7) or (8). For $\hat{\Delta}_{IPW3}$, ψ_{Δ} implied by (12) also depends on η_0, η_1, and this equation is solved jointly with those in (11) and (3); similarly ψ_{Δ} corresponding to $\hat{\Delta}_{DR}$ in (9) depends on $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1$, which are estimated by solving equations of the form $\sum_{i=1}^{n} I(Z_i = z)\psi_{\alpha_z}(Y_i, \mathbf{X}_i, \boldsymbol{\alpha}_z) = \mathbf{0}$, $z = 0, 1$, as for OLS or logistic regression.

This representation allows application of the theory of M-estimation; a review is given by Stefanski and Boos [18]. From Equation (3) of Reference [18], because the expectations of ψ_{β}, ψ_{η}, and ψ_{Δ} for $\hat{\Delta}_{IPW1}$, $\hat{\Delta}_{IPW2}$, and $\hat{\Delta}_{IPW3}$ are zero at the true values of $\boldsymbol{\beta}$, η_0, η_1, and Δ, the estimators of these quantities converge in probability to the true values, and hence, $\hat{\Delta}_{IPW1}$, $\hat{\Delta}_{IPW2}$, and $\hat{\Delta}_{IPW3}$ are consistent for Δ_0, the true value of Δ. (This may be seen equivalently by substituting the true values of $\boldsymbol{\beta}$, η_0, and η_1 in (7), (8), and (12) and applying the law of large numbers directly.) A similar argument shows that $\hat{\Delta}_{DR}$ converges in probability to Δ_0, even if the models m_z are not correctly specified, as the corresponding ψ_{Δ} still has mean zero. The theory [18, Section 2] then implies that each estimator is such that $n^{1/2}(\hat{\Delta} - \Delta_0)$ converges in distribution to a $N(0, \Sigma)$ random variable.

It instructive to first consider the (unlikely) case where $\boldsymbol{\beta}$ is known, so that $e(\mathbf{X}, \boldsymbol{\beta})$ is a known function of $\mathbf{X}$ and joint solution with (3) is unnecessary. Under these conditions, for $\hat{\Delta}_{IPW1}, \hat{\Delta}_{IPW2}$, and $\hat{\Delta}_{IPW3}$, the large-sample variances are

$$\begin{aligned} \Sigma^*_{IPW1} &= E\left(\frac{Y_1^2}{e} + \frac{Y_0^2}{1-e}\right) - \Delta_0^2, \quad \Sigma^*_{IPW2} = E\left\{\frac{(Y_1-\mu_1)^2}{e} + \frac{(Y_0-\mu_0)^2}{1-e}\right\} \\ \Sigma^*_{IPW3} &= E\left\{\frac{(Y_1-\mu_1)^2}{e} + \frac{(Y_0-\mu_0)^2}{1-e}\right\} + \eta_1 E\left(\frac{Y_1-\mu_1}{e}\right) + \eta_0 E\left(\frac{Y_0-\mu_0}{1-e}\right) + 2\eta_1\eta_0 \end{aligned} \tag{13}$$

where expectations are with respect to the distribution of $(Y_0, Y_1, \mathbf{X})$ and all parameters are equal to their true values. It may be shown that $\Sigma^*_{IPW2} \geqslant \Sigma^*_{IPW3}$. If, as in practice, $\boldsymbol{\beta}$ is estimated, then the variances become, with $\mathbf{E}_{\beta\beta} = E[e_{\beta} e_{\beta}^{\mathrm{T}}/\{e(1-e)\}]$,

$$\Sigma_{IPW1} = \Sigma^*_{IPW1} - \mathbf{H}_{\beta,1}^{\mathrm{T}} \mathbf{E}_{\beta\beta}^{-1} \mathbf{H}_{\beta,1}, \quad \mathbf{H}_{\beta,1} = E\left\{\left(\frac{Y_1}{e} + \frac{Y_0}{1-e}\right) e_{\beta}\right\} \tag{14}$$

$$\Sigma_{IPW2} = \Sigma^*_{IPW2} - \mathbf{H}_{\beta,2}^{\mathrm{T}} \mathbf{E}_{\beta\beta}^{-1} \mathbf{H}_{\beta,2}, \quad \mathbf{H}_{\beta,2} = E\left\{\left(\frac{Y_1-\mu_1}{e} + \frac{Y_0-\mu_0}{1-e}\right) e_{\beta}\right\} \tag{15}$$

$$\Sigma_{IPW3} = \Sigma^*_{IPW3} - \mathbf{H}_{\beta,3}^{\mathrm{T}} \mathbf{E}_{\beta\beta}^{-1} \mathbf{H}_{\beta,3}, \quad \mathbf{H}_{\beta,3} = E\left\{\left(\frac{Y_1-\mu_1+\eta_1}{e} + \frac{Y_0-\mu_0+\eta_0}{1-e}\right) e_{\beta}\right\} \tag{16}$$

thus exhibiting the interesting property that estimating $\boldsymbol{\beta}$, even if its true value is known, leads to smaller (large-sample) variance for these estimators than using the true value. Thus,

even if the functional form of the propensity score is known exactly, it is beneficial from an efficiency standpoint to estimate it anyway. We have found in empirical studies like those in Section 4 that in general $\Sigma_{IPW1} \geqslant \Sigma_{IPW2} \geqslant \Sigma_{IPW3}$.

For $\hat{\Delta}_{DR}$, similar arguments show that its large-sample variance is

$$\Sigma_{DR} = \Sigma^*_{IPW2} - E\left[\sqrt{\frac{1-e}{e}}\{E(Y_1 \mid \mathbf{X}) - \mu_1\} + \sqrt{\frac{e}{1-e}}\{E(Y_0 \mid \mathbf{X}) - \mu_0\}\right]^2 \tag{17}$$

The Robins *et al.* [13] theory guarantees that $\Sigma_{DR} \leqslant \Sigma_{IPW1}$, Σ_{IPW2}, and Σ_{IPW3}. As long as the propensity and regression models do not share parameters, Σ_{DR} is the same whether $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1$ are known or estimated.

The components of the expressions in (14)–(17) may be estimated from the observed data, yielding approximate sampling variances for $\hat{\Delta}_{IPW1}$, $\hat{\Delta}_{IPW2}$, $\hat{\Delta}_{IPW3}$, and $\hat{\Delta}_{DR}$. Alternatively, variance estimates may be obtained via the empirical sandwich method [18, Sections 2 and 3], which we have found to be more stable in practice. Specifically, for propensity models of the form $\{1+\exp(-\mathbf{W}^{\mathrm{T}}\boldsymbol{\beta})\}^{-1}$, where $\mathbf{W}$ is a function of elements in $\mathbf{X}$, approximate sampling variances are computed as $n^{-2}\sum_{i=1}^n \hat{I}_i^2$, where

$$\hat{I}_{IPW1,i} = \frac{Z_i Y_i}{\hat{e}_i} - \frac{(1-Z_i)Y_i}{1-\hat{e}_i} - \hat{\Delta}_{IPW1} - (Z_i - \hat{e}_i)\hat{\mathbf{H}}_{\beta,1}^{\mathrm{T}}\hat{\mathbf{E}}_{\beta\beta}^{-1}\mathbf{W}_i \tag{18}$$

$$\hat{I}_{IPW2,i} = \frac{Z_i(Y_i - \hat{\mu}_{1,IPW2})}{\hat{e}_i} - \frac{(1-Z_i)(Y_i - \hat{\mu}_{0,IPW2})}{1-\hat{e}_i} - (Z_i - \hat{e}_i)\hat{\mathbf{H}}_{\beta,2}^{\mathrm{T}}\hat{\mathbf{E}}_{\beta\beta}^{-1}\mathbf{W}_i \tag{19}$$

$$\begin{aligned}\hat{I}_{IPW3,i} = {} & \frac{Z_i(Y_i - \hat{\mu}_{1,IPW3}) + \hat{\eta}_1(Z_i - \hat{e}_i)}{\hat{e}_i} - \frac{(1-Z_i)(Y_i - \hat{\mu}_{0,IPW3}) - \hat{\eta}_0(Z_i - \hat{e}_i)}{1-\hat{e}_i} \\ & -(Z_i - \hat{e}_i)\hat{\mathbf{H}}_{\beta,3}^{\mathrm{T}}\hat{\mathbf{E}}_{\beta\beta}^{-1}\mathbf{W}_i\end{aligned} \tag{20}$$

$$\hat{I}_{DR,i} = \frac{Z_i Y_i - m_1(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_1)(Z_i - \hat{e}_i)}{\hat{e}_i} - \frac{(1-Z_i)Y_i + m_0(\mathbf{X}_i, \hat{\boldsymbol{\alpha}}_0)(Z_i - \hat{e}_i)}{(1-\hat{e}_i)} - \hat{\Delta}_{DR} \tag{21}$$

$\hat{\mathbf{E}}_{\beta\beta}^{-1} = n^{-1}\sum_{i=1}^n \hat{e}_i(1-\hat{e}_i)\mathbf{W}_i\mathbf{W}_i^{\mathrm{T}}$, $\hat{\eta}_1 = -\sum_{i=1}^n \{Z_i(Y_i - \hat{\mu}_{1,IPW3})/\hat{e}_i^2\} \big/ \sum_{i=1}^n \{(Z_i - \hat{e}_i)/\hat{e}_i\}^2$, and $\hat{\eta}_0 = -\sum_{i=1}^n \{(1-Z_i)(Y_i - \hat{\mu}_{0,IPW3})/(1-\hat{e}_i)^2\} \big/ \sum_{i=1}^n \{(Z_i - \hat{e}_i)/(1-\hat{e}_i)\}^2$. The terms $\hat{\mathbf{H}}_{\beta,1}$, $\hat{\mathbf{H}}_{\beta,2}$, and $\hat{\mathbf{H}}_{\beta,3}$ are empirical versions of the terms in (14)–(16):

$$\hat{\mathbf{H}}_{\beta,1} = n^{-1}\sum_{i=1}^n \left\{\frac{Z_i Y_i(1-\hat{e}_i)}{\hat{e}_i} + \frac{(1-Z_i)Y_i\hat{e}_i}{1-\hat{e}_i}\right\}\mathbf{W}_i$$

$$\hat{\mathbf{H}}_{\beta,2} = n^{-1}\sum_{i=1}^n \left\{\frac{Z_i(Y_i - \hat{\mu}_{1,IPW2})(1-\hat{e}_i)}{\hat{e}_i} + \frac{(1-Z_i)(Y_i - \hat{\mu}_{0,IPW2})\hat{e}_i}{1-\hat{e}_i}\right\}\mathbf{W}_i$$

$$\hat{\mathbf{H}}_{\beta,3} = n^{-1}\sum_{i=1}^n \left\{\frac{Z_i(Y_i - \hat{\mu}_{1,IPW3} + \hat{\eta}_1)(1-\hat{e}_i)}{\hat{e}_i} + \frac{(1-Z_i)(Y_i - \hat{\mu}_{0,IPW3} + \hat{\eta}_0)\hat{e}_i}{1-\hat{e}_i}\right\}\mathbf{W}_i$$

In Section 4, we demonstrate performance of these formulæ.

3.2. Stratification estimators

Here, we present a heuristic account of large-sample results for $\hat{\Delta}_S$ and $\hat{\Delta}_{SR}$ based on representing the stratification and within-stratum estimation schemes for each as solutions to sets of estimating equations. Because in practice it is standard to take a predetermined number of strata K regardless of sample size ($K=5$ is most common), we view K as fixed (so not depending on n). As in Section 3.1, assume e is correctly specified.

Both $\hat{\Delta}_S$ and $\hat{\Delta}_{SR}$ involve estimation not only of $\boldsymbol{\beta}$ by solving (3), as before, but also of the true quantiles $\mathbf{q}=(q_1,\ldots,q_{K-1})^{\mathrm{T}}$, which may be carried out by solving

$$\sum_{i=1}^{n}\psi^{S}_{q_j}(\mathbf{X}_i,q_j,\boldsymbol{\beta})=\sum_{i=1}^{n}I(e_i\leqslant q_j)-j/K=0,\quad j=1,\ldots,K-1 \tag{22}$$

These equations do not have zero solutions for some n, but this technicality does not affect the spirit of the discussion below. We may rewrite (4) in an asymptotically equivalent form by replacing n_j/n with its limit K^{-1} and writing $\hat{p}_j=n_{1j}/n$ as

$$\hat{\Delta}_S=n^{-1}\sum_{i=1}^{n}\frac{Z_iY_i}{K}\left\{\sum_{j=1}^{K}\frac{I(\hat{e}_i\in\hat{Q}_j)}{\hat{p}_j}\right\}-n^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{K}\left\{\sum_{j=1}^{K}\frac{I(\hat{e}_i\in\hat{Q}_j)}{1/K-\hat{p}_j}\right\} \tag{23}$$

This shows that $\hat{\Delta}_S$ also requires estimation of the probabilities $\mathbf{p}=(p_1,\ldots,p_K)^{\mathrm{T}}$ that an individual is treated and has propensity score in $Q_j=(q_{j-1},q_j]$, where $q_0=0$, $q_K=1$; the estimator $\hat{p}_j=n_{1_j}/n$ follows from solving the equations

$$\sum_{i=1}^{n}\psi^{S}_{p_j}(Z_i,\mathbf{X}_i,q_{j-1},q_j,p_j,\boldsymbol{\beta})=\sum_{i=1}^{n}Z_iI(e_i\in Q_j)-p_j=0,\quad j=1,\ldots,K \tag{24}$$

Instead, calculation of $\hat{\Delta}_{SR}$ involves solving in $\boldsymbol{\alpha}^{(j)}$ for $j=1,\ldots,K$

$$\sum_{i=1}^{n}\psi^{S}_{\alpha^{(j)}}(Y_i,Z_i,\mathbf{X}_i,q_{j-1},q_j,\boldsymbol{\alpha}^{(j)})=\sum_{i=1}^{n}I(e_i\in Q_j)\{Y_i-m^{(j)}(Z_i,\mathbf{X}_i,\boldsymbol{\alpha}^{(j)})\}m^{(j)}_{\alpha}(Z_i,\mathbf{X}_i,\boldsymbol{\alpha}^{(j)})=\mathbf{0} \tag{25}$$

where $m^{(j)}_{\alpha}$ is the vector of partial derivatives of $m^{(j)}$ with respect to elements of $\boldsymbol{\alpha}^{(j)}$. We are now in a position to characterize fully each estimator and evaluate properties.

First consider $\hat{\Delta}_S$. Even with $e(\mathbf{X},\boldsymbol{\beta})$ correctly specified, as noted in Section 2.3, we expect $\hat{\Delta}_S$ to be inconsistent due to failure of stratification to eliminate all confounding, an observation we may now formalize. Noting that (3), (22), and (24) have expectation zero at the true values of $\boldsymbol{\theta}=(\mathbf{q}^{\mathrm{T}},\mathbf{p}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, we may conclude from [18, Section 2] that solving these equations jointly yields consistent estimators for the elements of $\boldsymbol{\theta}$. Thus, considering the asymptotically equivalent form (23), we may replace $\hat{e}_i$, $\hat{Q}_j$, and $\hat{p}_j$ by their true values and apply the law of large numbers directly to see that $\hat{\Delta}_S$ converges in probability to $\Delta_*=\mu_1^*-\mu_0^*$, where $\mu_1^*=K^{-1}\sum_{j=1}^{K}E\{Y_1eI(e\in Q_j)\}/E\{eI(e\in Q_j)\}$, and $\mu_0^*=K^{-1}\sum_{j=1}^{K}E\{Y_0(1-e)I(e\in Q_j)\}/[K^{-1}-E\{eI(e\in Q_j)\}]$. It is straightforward to see that a sufficient condition for $\Delta_*=\Delta_0$ is $(Y_0,Y_1)\perp\!\!\!\perp\mathbf{X}$, in which case confounding is not an issue, as would be expected, but, in general, $\Delta_*\neq\Delta_0$ so that $\hat{\Delta}_S$ is not consistent. The hope in practice, of course, is that

$|\Delta_* - \Delta_0|$ is 'small.' Thus, $\hat{\Delta}_S$ estimates Δ_*, and from (23) an estimating equation for Δ_* is $\sum_{i=1}^n \psi^S_{\Delta_*}(Y_i, Z_i, \mathbf{X}, \Delta^*, \boldsymbol{\theta}) = 0$, where

$$\psi^S_{\Delta_*}(Y_i, Z_i, \mathbf{X}, \Delta^*, \boldsymbol{\theta}) = Z_i Y_i K^{-1} \sum_{j=1}^K I(e_i \in Q_j)/p_j - (1 - Z_i) Y_i K^{-1}$$

$$\times \sum_{j=1}^K I(e_i \in Q_j)/(K^{-1} - p_j) - \Delta_*$$

Writing $\boldsymbol{\Psi}_\theta = (\psi^S_{q_1}, \ldots, \psi^S_{q_{K-1}}, \psi^S_{p_1}, \ldots, \psi^S_{p_K}, \psi_\beta)^{\mathrm{T}}$, we thus see that $\hat{\Delta}_S$ and $\hat{\boldsymbol{\theta}}$ jointly solve

$$\sum_{i=1}^n \{\boldsymbol{\Psi}^{\mathrm{T}}_\theta(Z_i, \mathbf{X}_i, \boldsymbol{\theta}), \psi^S_{\Delta^*}(Y_i, Z_i, \mathbf{X}, \Delta_*, \boldsymbol{\theta})\}^{\mathrm{T}} = \mathbf{0} \tag{26}$$

in $\boldsymbol{\theta}$ and Δ_*. The properties of $\hat{\Delta}_S$ may be derived from (26) by appealing to M-estimation arguments [18]. Consider first the 'ideal' situation where the q_j, p_j, and $\boldsymbol{\beta}$ are all known. Letting $f_e(\cdot)$ be the density of the propensity score and $E(\cdot \mid e)$ be conditional expectation given the propensity score, it may be shown under these conditions that $n^{1/2}(\hat{\Delta}_S - \Delta_*)$ converges in distribution to a $N(0, \Sigma^*_S)$ random variable, where

$$\Sigma^*_S = K^{-2} \sum_{j=1}^K p_j^{-2} \int_{q_{j-1}}^{q_j} E(Y_1^2 \mid t) t f_e(t)\, \mathrm{d}t + K^{-2} \sum_{j=1}^K (1/K - p_j)^{-2}$$

$$\times \int_{q_{j-1}}^{q_j} E(Y_0^2 \mid t)(1 - t) f_e(t)\, \mathrm{d}t - \Delta_*^2$$

Comparing this expression to those in (13) suggests that $\hat{\Delta}_S$ has different properties from weighted estimators, as Σ^*_S depends critically on the density of the propensity score. In the more realistic case where the q_j, p_j, and $\boldsymbol{\beta}$ are estimated, via M-estimation arguments for nonsmooth ψ functions [18, Section 4] to account for nondifferentiability of some elements of (26) in q_j and $\boldsymbol{\beta}$, the variance is

$$\Sigma_S = \Sigma^*_S + \boldsymbol{\Gamma}_p + \boldsymbol{\Gamma}_{qp} + \boldsymbol{\Gamma}_{\beta qp} \tag{27}$$

where $\boldsymbol{\Gamma}_p$, $\boldsymbol{\Gamma}_{qp}$, and $\boldsymbol{\Gamma}_{\beta qp}$ are quantities modifying the 'ideal' variance Σ^* due to estimation in turn of $\mathbf{p}$, $\mathbf{q}$, and $\boldsymbol{\beta}$; e.g. $\boldsymbol{\Gamma}_{\beta qp}$ is the effect of estimating $\boldsymbol{\beta}$ rather than knowing it if $\mathbf{q}$ and $\mathbf{p}$ are estimated (see the Appendix). In contrast to the situation in (14), (15), and (16), it is not possible to deduce that any of $\boldsymbol{\Gamma}_p$, $\boldsymbol{\Gamma}_{qp}$, or $\boldsymbol{\Gamma}_{\beta qp}$ in (27) are negative, which would imply that estimation of $\mathbf{p}$, $\mathbf{q}$, and/or $\boldsymbol{\beta}$ reduces variance relative to the (unlikely) situation where they are known.

We may follow a similar argument for $\hat{\Delta}_{SR}$. This estimator requires joint solution of (3), (22), and (25); as above, solving the first two jointly leads to consistent estimators for $\boldsymbol{\beta}$ and the q_j. Substituting these in (25), from the theory of M-estimation [18, Section 2], the resulting estimators $\hat{\boldsymbol{\alpha}}^{(j)}$, $j = 1, \ldots, K$, solving (25) converge in probability to some $\boldsymbol{\alpha}^{(j)}_*$

satisfying $E\{\psi^S_{\alpha^{(j)}}(Y,Z,\mathbf{X},q_{j-1},q_j,\boldsymbol{\alpha}^{(j)}_*)\}=\mathbf{0}$ for each j, where $\boldsymbol{\alpha}^{(j)}_*$ depend on the functions $m^{(j)}$ used. Now, substituting $n_j/n\approx K^{-1}$ in (5), we may rewrite (6) as

$$\hat{\Delta}_{SR}=n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{K}I(\hat{e}_i\in\hat{Q}_j)\{m^{(j)}(1,\mathbf{X}_i,\hat{\boldsymbol{\alpha}}^{(j)})-m^{(j)}(0,\mathbf{X}_i,\hat{\boldsymbol{\alpha}}^{(j)})\} \tag{28}$$

Then, applying the law of large numbers to (28), $\hat{\Delta}_{SR}$ converges in probability to $\Delta_{**}=\sum_{j=1}^{K}E[I(e\in Q_j)\{m^{(j)}(1,\mathbf{X},\boldsymbol{\alpha}^{(j)}_*)-m^{(j)}(0,\mathbf{X},\boldsymbol{\alpha}^{(j)}_*)\}]$; e.g. for the linear model example following (5), $\Delta_{**}=\sum_{j=1}^{K}E\{I(e\in Q_j)\}\boldsymbol{\alpha}^{(j)}_*=K^{-1}\sum_{j=1}^{K}\boldsymbol{\alpha}^{(j)}_*$. If the within-stratum regression models $m^{(j)}(Z,\mathbf{X},\boldsymbol{\alpha}^{(j)})$ are chosen such that they are all of the exact form of the true regression relationship $E(Y\mid Z,\mathbf{X})=m(Z,\mathbf{X},\boldsymbol{\alpha}_0)$, say, for some m and true value $\boldsymbol{\alpha}_0$, then $\boldsymbol{\alpha}^{(j)}_*=\boldsymbol{\alpha}_0$ for each j, as under these conditions $E\{\psi^S_{\alpha^{(j)}}(Y,Z,\mathbf{X},q_{j-1},q_j,\boldsymbol{\alpha}_0)\}=E(I(e\in Q_j)E[\{Y-m(Z,\mathbf{X},\boldsymbol{\alpha}_0)\}\mid Z,\mathbf{X}]\, m_\alpha(Z,\mathbf{X},\boldsymbol{\alpha}_0))=\mathbf{0}$ because the inner conditional expectation is zero. Thus, using (2) and $m(z,\mathbf{X},\boldsymbol{\alpha}_0)=E(Y\mid Z=z,\mathbf{X})$

$$\begin{aligned}\Delta_{**}&=\sum_{j=1}^{K}E[I(e\in Q_j)\{E(Y\mid Z=1,\mathbf{X})-E(Y\mid Z=0,\mathbf{X})\}]\\&=E\left[\left\{\sum_{j=1}^{K}I(e\in Q_j)\right\}\{E(Y_1\mid\mathbf{X})-E(Y_0\mid\mathbf{X})\}\right]=E\{E(Y_1\mid\mathbf{X})-E(Y_0\mid\mathbf{X})\}=\Delta_0\end{aligned}$$

where we use the facts that the sum over j of the indicators of stratum membership is one for any fixed $\mathbf{X}$ and $E\{E(Y_1\mid\mathbf{X})-E(Y_0\mid\mathbf{X})\}$ is equal to the true value of Δ. This demonstrates that $\hat{\Delta}_{SR}$ is a consistent estimator for Δ_0 as long as the $m^{(j)}$ have the same form as the true regression relationship. However, if the $m^{(j)}$ are chosen differently, and hence incorrectly, this argument does not hold, and $\Delta_{**}\neq\Delta_0$ in general. Hence, choice of the within-stratum regression models is critical for consistency of $\hat{\Delta}_{SR}$. In contrast, by 'double robustness,' Δ_{DR}, will be consistent regardless of whether the regression models chosen for 'augmentation' are correct. In Section 4, we demonstrate these properties empirically.

Analogous to the results for $\hat{\Delta}_S$, again by the theory of M-estimation, it may be shown that in general $n^{1/2}(\hat{\Delta}_{SR}-\Delta_{**})$ converges in distribution to a normal random variable with variance similar in form to that in (27); thus, no general insights are possible.

Such theory is not used in practice; rather, it is routine to approximate the sampling variance of $\hat{\Delta}_S$ by treating $\hat{\Delta}_S$ as the average of K independent, within-stratum, treatment effect estimates as

$$K^{-2}\sum_{j=1}^{K}\hat{\sigma}^2_j \tag{29}$$

assuming an equal number of individuals per stratum, where $\hat{\sigma}^2_j$ is an estimate of the variance of the difference between the treatment means in stratum j given by $\hat{\sigma}^2_j=n_{1j}^{-1}s^2_{1j}+(n_j-n_{1j})^{-1}s^2_{0j}$, $s^2_{1j}=n_{1j}^{-1}\sum_{i=1}^{n}I(\hat{e}_i\in\hat{Q}_j)(Z_iY_i-\bar{y}_{1j})^2$, $s^2_{0j}=(n_j-n_{1j})^{-1}\sum_{i=1}^{n}I(\hat{e}_i\in\hat{Q}_j)\{(1-Z_i)Y_i-\bar{y}_{0j}\}^2$, $\bar{y}_{1j}=n_{1j}^{-1}\sum_{i=1}^{n}I(\hat{e}_i\in\hat{Q}_j)Z_iY_i$, and $\bar{y}_{0j}=(n_j-n_{1j})^{-1}\sum_{i=1}^{n}I(\hat{e}_i\in\hat{Q}_j)(1-Z_i)Y_i$. Similarly,

the sampling variance of $\hat{\Delta}_{SR}$ is approximated in practice by an expression of the form (29) with $\hat{\sigma}_j^2$ replaced by an estimate of the variance of $m^{(j)}(1, \mathbf{X}, \hat{\boldsymbol{\alpha}}^{(j)}) - m^{(j)}(0, \mathbf{X}, \hat{\boldsymbol{\alpha}}^{(j)})$ based on the fit of the regression model in stratum j; e.g., for the linear model example after (6), this would be the estimated sampling variance of $\hat{\alpha}_Z^{(j)}$, obtainable directly from standard regression software.

3.3. *Effect of additional covariates*

In the previous development, it was assumed that $\mathbf{X}$ is associated with both treatment exposure and potential response and that (2) holds. For $\hat{\Delta}_S$, a common guideline is that it is preferable to 'over-model' the propensity score by including additional covariates unrelated to treatment exposure rather than run the risk of excluding relevant ones [5, 19]. In fact, intuition would suggest that including such covariates when they are correlated with potential response could provide additional information on Δ. It is possible to gain formal insight as follows.

Suppose $\mathbf{V}$ is an additional set of covariates, exclusive of $\mathbf{X}$, that (i) is not associated with treatment exposure but (ii) is associated with potential response. More precisely, (i) may be written as $P(Z = 1 \mid \mathbf{X}, \mathbf{V}) = P(Z = 1 \mid \mathbf{X})$, and (ii) implies that the conditional distributions of Y_0 and Y_1 given $(Z, \mathbf{X}, \mathbf{V})$ depend on $\mathbf{V}$. Suppose that the analyst is willing to assume strong ignorability given both $\mathbf{X}$ and $\mathbf{V}$, i.e.

$$(Y_0, Y_1) \perp\!\!\!\perp Z \mid (\mathbf{X}, \mathbf{V}) \tag{30}$$

It is straightforward to show using manipulations similar to those in Reference [20] that here (30) implies that (2) also holds. Thus, it is possible to specify a model $P(Z = 1 \mid \mathbf{X}, \mathbf{V}) = e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is an additional $(q \times 1)$ parameter corresponding to terms in the model involving $\mathbf{V}$, such that this model reduces to the true propensity score $e(\mathbf{X}, \boldsymbol{\beta})$ (depending on $\mathbf{X}$ and $\boldsymbol{\beta}$ only) when $\boldsymbol{\gamma} = \mathbf{0}$, its 'true' value, and the assumptions underlying the derivations of (14)–(17) and (27) hold. Suppose, then, that the chosen propensity score model satisfies $e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \mathbf{0}) = e(\mathbf{X}, \boldsymbol{\beta}) = e$ and is such that $\partial/\partial\boldsymbol{\beta}\{e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\}|_{\gamma=0} = e_\beta$ depending on $\mathbf{X}$ and $\boldsymbol{\beta}$ only; e.g. as for the logistic model $e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = [1 + \exp\{-(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{V}^{\mathrm{T}}\boldsymbol{\gamma})\}]^{-1}$.

Under these circumstances, for all methods, Δ will be estimated jointly with both the previous additional parameters and $\boldsymbol{\gamma}$. The effect of including $\mathbf{V}$ in the propensity model may thus be deduced by considering the previous estimating equations for each estimator, replacing $e(\mathbf{X}, \boldsymbol{\beta})$ by $e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, and adding the additional equation

$$\sum_{i=1}^{n} \frac{Z_i - e(\mathbf{X}_i, \mathbf{V}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})}{e(\mathbf{X}_i, \mathbf{V}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\{1 - e(\mathbf{X}_i, \mathbf{V}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\}} \partial/\partial\boldsymbol{\gamma}\{e(\mathbf{X}_i, \mathbf{V}_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\} = \mathbf{0} \tag{31}$$

Note that $\partial/\partial\boldsymbol{\gamma}\{e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\}$ evaluated at the 'truth' $\boldsymbol{\gamma} = \mathbf{0}$ may depend on both $\mathbf{X}$ and $\mathbf{V}$; in the logistic example, this partial derivative is $\mathbf{V}/[e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\{1 - e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\}]$. In general, write $e_\gamma = \partial/\partial\boldsymbol{\gamma}\{e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\gamma})\}|_{\gamma=0}$, with subscript i when evaluated at $(\mathbf{X}_i, \mathbf{V}_i)$.

Incorporating the additional estimating equation (31) for each estimator, it may be shown by M-estimation arguments [18] that all weighted estimators still are consistent and such that $n^{1/2}(\hat{\Delta} - \Delta_0)$ converges in distribution to a mean-zero normal random variable, now with different variance Σ^V. Defining $\mathbf{E}_{\gamma\gamma} = E[e_\gamma e_\gamma^{\mathrm{T}}/\{e(1-e)\}]$ and $\mathbf{E}_{\gamma\beta} = E[e_\gamma e_\beta^{\mathrm{T}}/\{e(1-e)\}]$, and

letting $\mathbf{H}_{\gamma\beta} = \mathbf{E}_{\gamma\gamma} - \mathbf{E}_{\gamma\beta}\mathbf{E}_{\beta\beta}^{-1}\mathbf{E}_{\gamma\beta}^{\mathrm{T}}$, for $\hat{\Delta}_{IPW2}$,

$$\Sigma_{IPW2}^{V} = \Sigma_{IPW2} - (\mathbf{H}_{\gamma,2} - \mathbf{E}_{\gamma\beta}\mathbf{E}_{\beta\beta}^{-1}\mathbf{H}_{\beta,2})^{\mathrm{T}}\mathbf{H}_{\gamma\beta}^{-1}(\mathbf{H}_{\gamma,2} - \mathbf{E}_{\gamma\beta}\mathbf{E}_{\beta\beta}^{-1}\mathbf{H}_{\beta,2}) \tag{32}$$

where $\mathbf{H}_{\gamma,2} = E[\{(Y_1 - \mu_1)/e + (Y_0 - \mu_0)/(1-e)\}e_{\gamma}]$, with similar expressions for Σ_{IPW1}^{V} and Σ_{IPW3}^{V}. From (32) and these analogous expressions, the effect of including $\mathbf{V}$ in the propensity score model is to reduce the variance relative to that in the case where $\mathbf{V}$ is excluded. The practical implication is that, at least in large samples, for these weighted estimators, incorporating covariates in the propensity model that are not related to treatment exposure but are associated with potential response will always lead to precision for estimating Δ at least as great as that attained by disregarding such covariates.

When $\mathbf{V}$ is considered, the form of the semiparametric efficient estimator, which now is that with smallest large-sample variance among all estimators in the Robins *et al.* [13] class under the condition that the distribution of $(Y_0, Y_1, \mathbf{X}, \mathbf{V})$ is unspecified, is different from (9), which does not acknowledge availability of $\mathbf{V}$. In particular, we now have

$$\hat{\Delta}_{DR} = n^{-1}\sum_{i=1}^{n}\frac{Z_iY_i - (Z_i - \hat{\hat{e}}_i)m_1^*(\mathbf{X}_i, \mathbf{V}_i, \hat{\boldsymbol{\delta}}_1)}{\hat{\hat{e}}_i} - n^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i + (Z_i - \hat{\hat{e}}_i)m_0^*(\mathbf{X}_i, \mathbf{V}_i, \hat{\boldsymbol{\delta}}_0)}{1 - \hat{\hat{e}}_i} \tag{33}$$

where $\hat{\hat{e}}_i = e(\mathbf{X}_i, \mathbf{V}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, and $m_z^*(\mathbf{X}, \mathbf{V}, \boldsymbol{\delta}_z) = E(Y \mid Z = z, \mathbf{X}, \mathbf{V})$ is the regression of Y on $(\mathbf{X}, \mathbf{V})$ in group $z, z = 0, 1$, depending on parameters $\boldsymbol{\delta}_z$ estimated by $\hat{\boldsymbol{\delta}}_z$ from subjects with $Z = z$. As before, this estimator requires modelling of the regression and maintains the 'double-robustness' property. The large sample variance of (33)

$$\Sigma_{DR}^{V} = \Sigma_{IPW2}^{*} - E\left[\sqrt{\frac{1-e}{e}}\{E(Y_1 \mid \mathbf{X}, \mathbf{V}) - \mu_1\} + \sqrt{\frac{e}{1-e}}\{E(Y_0 \mid \mathbf{X}, \mathbf{V}) - \mu_0\}\right]^2$$

and satisfies $\Sigma_{DR}^{V} \leqslant \Sigma_{DR}$, so that a potential gain in efficiency over disregarding information on Y in $\mathbf{V}$ is achieved. Of course, $\Sigma_{DR}^{V} \leqslant \Sigma_{IPW1}^{V}, \Sigma_{IPW2}^{V}$, and Σ_{IPW3}^{V} as well.

As $e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \mathbf{0}) = e$, $\hat{\Delta}_S$ and $\hat{\Delta}_{SR}$ still converge in probability to Δ_* and Δ_{**} in general; however, the large-sample variances change. For example, for $\hat{\Delta}_S$, by similar arguments, where now (31) is solved jointly with the previous equations, the variance is (see the appendix)

$$\Sigma_S^{V} = \Sigma_S + \boldsymbol{\Gamma}_{\gamma\beta qp} \tag{34}$$

where $\boldsymbol{\Gamma}_{\gamma\beta qp}$ represents the additional effect of estimating $\boldsymbol{\gamma}$ rather than knowing it if $(\mathbf{p}, \mathbf{q}, \boldsymbol{\beta})$ are estimated; as before, it is not possible to show $\boldsymbol{\Gamma}_{\gamma\beta qp} \leqslant 0$. A similar development holds for $\hat{\Delta}_{SR}$, where we still have $\Delta_{**} = \Delta_0$ if the $m^{(j)}$ are chosen according to the true regression relationship. Thus, in contrast to the results for weighted estimators, it is not immediately evident whether incorporating covariates $\mathbf{V}$ into the propensity model leads to a reduction in variance for these estimators over not. In Section 4, we investigate this issue empirically.

4. SIMULATION STUDIES

In practice, several covariates will likely be available for modelling the propensity score. To investigate relative performance in such a realistic setting, we carried out simulations involving a number of continuous and discrete covariates and a continuous response such that $\Delta_0 > 0$, where larger values of the response are preferred, so that treatment is beneficial.

We considered covariates $\mathbf{X} = (X_1, X_2, X_3)^{\mathrm{T}}$ associated with both treatment exposure and potential response, i.e. confounders, and covariates $\mathbf{V} = (V_1, V_2, V_3)^{\mathrm{T}}$ associated with potential response but not treatment exposure, so that the effect of adding such covariates as in Section 3.3 could be gauged. In particular, in all scenarios, Z was generated as Bernoulli according to the true propensity score $e(\mathbf{X}, \boldsymbol{\beta}) = \{1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3)\}^{-1}$, not involving elements of $\mathbf{V}$, and the response Y was generated according to

$$Y = v_0 + v_1 X_1 + v_2 X_2 + v_3 X_3 + v_4 Z + \xi_1 V_1 + \xi_2 V_2 + \xi_3 V_3 + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0,1) \tag{35}$$

and $\mathbf{v} = (v_0, v_1, v_2, v_3, v_4)^{\mathrm{T}} = (0, -1, 1, -1, 2)^{\mathrm{T}}$, so that in all cases $\Delta_0 = v_4 = 2$. Settings of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ and $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^{\mathrm{T}}$ were chosen to represent different degrees of association, as described below. All scenarios are such that values of $\mathbf{X}$ associated with lower responses are also associated with increased propensity for treatment, so that subjects with a covariate profile indicating poor response are those more likely to be treated.

The joint distribution of $(\mathbf{X}, \mathbf{V})$ was specified by taking $X_3 \sim$ Bernoulli (0.2) and then generating V_3 as Bernoulli with $P(V_3 = 1 \mid X_3) = 0.75 X_3 + 0.25(1 - X_3)$. Conditional on X_3, $(X_1, V_1, X_2, V_2)^{\mathrm{T}}$ was then generated as multivariate normal $\mathrm{N}(\boldsymbol{\tau}_{X_3}, \boldsymbol{\Sigma}_{X_3})$, where

$$\boldsymbol{\tau}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \quad \boldsymbol{\tau}_0 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & 0.5 & -0.5 & -0.5 \\ 0.5 & 1 & -0.5 & -0.5 \\ -0.5 & -0.5 & 1 & 0.5 \\ -0.5 & -0.5 & 0.5 & 1 \end{pmatrix}$$

Values for $\mathbf{v}$ and $\boldsymbol{\xi}$ were taken such that each positively-correlated pair $(X_k, V_k), k = 1, 2$, has coefficients of the same sign in (35) and thus X_k and V_k have similar and correlated effects on response. Overall, the values for $\mathbf{v}, \boldsymbol{\beta}$, and $\boldsymbol{\xi}$ result in lower response values and larger probabilities of treatment exposure when $X_3 = 1$ and conversely when $X_3 = 0$. Note that (35) implies $E(Y \mid Z = z, \mathbf{X}, \mathbf{V}) = v_0 + v_1 X_1 + v_2 X_2 + v_3 X_3 + v_4 z + \xi_1 V_1 + \xi_2 V_2 + \xi_3 V_3 = m_z^*(\mathbf{X}, \mathbf{V}, \boldsymbol{\delta}_z)$ for $z = 0, 1$, where $\boldsymbol{\delta}_0 = (v_0, v_1, v_2, v_3, \xi_1, \xi_2, \xi_3)^{\mathrm{T}}$, $\boldsymbol{\delta}_1 = (v_0^*, v_1, v_2, v_3, \xi_1, \xi_2, \xi_3)^{\mathrm{T}}$, and $v_0^* = v_0 + v_4$. Moreover, this formulation implies expressions of the form $E(Y \mid Z = z, \mathbf{X}) = m_z(\mathbf{X}, \boldsymbol{\alpha}_z) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + v_4 z$ for some $\boldsymbol{\alpha}_0 = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}, \boldsymbol{\alpha}_1 = (\alpha_0^*, \alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$, and $\alpha_0^* = \alpha_0 + v_4$.

Settings of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ that achieve the features described above were chosen to represent varying degrees of association of the corresponding covariate to Z or Y. Three settings of $\boldsymbol{\xi}$ were used to examine the influence of the strength of the association between $\mathbf{V}$ and response when over-fitting the propensity score: $\boldsymbol{\xi}^{\mathrm{str}} = (-1.0, 1.0, 1.0)^{\mathrm{T}}, \boldsymbol{\xi}^{\mathrm{mod}} = (-0.5, 0.5, 0.5)^{\mathrm{T}}$, and $\boldsymbol{\xi}^{\mathrm{no}} = (0, 0, 0)^{\mathrm{T}}$, where superscripts no, mod, and str denote 'no,' 'moderate,' and 'strong' association. When $\boldsymbol{\xi} = \boldsymbol{\xi}^{\mathrm{no}}, \mathbf{V}$ is associated with neither potential response nor treatment exposure, so from Section 3.3 we expect no benefit to including it in an analysis. Two

settings $\boldsymbol{\beta}^{\text{str}} = (0.0, 0.6, -0.6, 0.6)^{\text{T}}$ and $\boldsymbol{\beta}^{\text{mod}} = (0.0, 0.3, -0.3, 0.3)^{\text{T}}$ were considered, corresponding to strong and moderate association of Z and $\mathbf{X}$, yielding marginal exposure probabilities $P(Z=1) = 0.38$ (str) and 0.42 (mod). For each of the six combinations of $(\boldsymbol{\xi}, \boldsymbol{\beta})$, 1000 Monte Carlo (MC) data sets were generated for $n = 1000$ and 5000 to emulate many published applications. For each, Δ was estimated using $\hat{\Delta}_{IPW1}, \hat{\Delta}_{IPW2}, \hat{\Delta}_{IPW3}, \hat{\Delta}_{DR}$, and $\hat{\Delta}_S$ and $\hat{\Delta}_{SR}$ with $K = 5$ two ways: (i) including only the true confounders $\mathbf{X}$ in the propensity score, as described in Sections 2.4 and 2.3, thus fitting the true propensity model $e(\mathbf{X}, \boldsymbol{\beta})$ above by ML, and (ii) including both $\mathbf{X}$ and $\mathbf{V}$ as described in Section 3.3, fitting the propensity model $e(\mathbf{X}, \mathbf{V}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \{1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - \gamma_1 V_1 - \gamma_2 V_2 - \gamma_3 V_3)\}^{-1}$ by ML. For $\hat{\Delta}_{DR}$, in (i), we fit the correct linear models $m_z(\mathbf{X}, \boldsymbol{\alpha}_z)$ implied above, and in (ii) we fit instead $m_z^*(\mathbf{X}, \mathbf{V}, \boldsymbol{\delta}_z)$, $z = 0, 1$, both by *OLS*. For $\hat{\Delta}_{SR}$, we similarly fit within each stratum the true models for $E(Y \mid Z, \mathbf{X})$ and $E(Y \mid Z, \mathbf{X}, \mathbf{V})$ for (i) and (ii), respectively. As discussed in Section 2.5, because *OLS* is *ML* estimation in this situation and hence serves as a 'benchmark,' we also estimated $\Delta_0 = v_4$ by directly fitting the true models for (i) $E(Y \mid Z, \mathbf{X})$ and (ii) $E(Y \mid Z, \mathbf{X}, \mathbf{V})$ by this method, denoted $\hat{\Delta}_{ML}$.

To investigate 'double robustness' of $\hat{\Delta}_{DR}$ and sensitivity of $\hat{\Delta}_{SR}$ and $\hat{\Delta}_{ML}$ to incorrect specification of regression models, for both (i) and (ii), we also implemented these estimators using the correct propensity models but mismodelling the relevant regression relationships by leaving (X_1, V_1) and X_1 out of the models for $E(Y \mid Z, \mathbf{X}, \mathbf{V})$ and $E(Y \mid Z, \mathbf{X})$, respectively, denoted by $\hat{\Delta}_{DR^*}$ and $\hat{\Delta}_{SR^*}$. Similarly, for $\hat{\Delta}_{ML}$, we fit these misspecified models directly by *OLS*, denoted by $\hat{\Delta}_{ML^*}$.

Table I summarizes results in the case where the regression models in $\hat{\Delta}_{DR}, \hat{\Delta}_{SR}$, and $\hat{\Delta}_{ML}$ correspond to the true relationships; as $\hat{\Delta}_{IPW1}$ performed uniformly more poorly than the other IPW estimators, it is omitted for brevity. Biases for all estimators but $\hat{\Delta}_S$ are less than 3 per cent in all scenarios, so are not shown. Those for $\hat{\Delta}_S$ under conditions (i) and (ii) can be substantial, particularly when associations are strong, demonstrating the inconsistency of this estimator. Thus, although MC standard deviation of $\hat{\Delta}_S$ is smaller than that of $\hat{\Delta}_{IPW2}$ and $\hat{\Delta}_{IPW3}$ in many cases, efficiency gains of the latter estimators over $\hat{\Delta}_S$ as measured by MC mean square error (MSE) ratio are considerable. In principle, in smaller sample sizes, biased estimators may outperform estimators with larger sampling variance, as the bias is small relative to the variance. However, in our experience, we have found this not to be true for $\hat{\Delta}_S$, with this estimator having bias far exceeding the bias $|\Delta_* - \Delta_0|$ predicted by the theory. The result is that weighted estimators achieve efficiency gains over $\hat{\Delta}_S$ at both small and large sample sizes, with comparable performance only in a limited range of moderate sample sizes (see Reference [21]). The estimator $\hat{\Delta}_{IPW3}$ has smaller variance than $\hat{\Delta}_{IPW2}$, particularly when $\boldsymbol{\beta} = \boldsymbol{\beta}^{\text{str}}$, showing that this estimator does indeed increase efficiency over simpler weighted estimators. However, the results for $\hat{\Delta}_{DR}$ and $\hat{\Delta}_{SR}$ shows that incorporation of regression modelling yields a substantial payoff. For the former, as predicted by the theory, MC standard deviations for these estimator are uniformly smaller than those for $\hat{\Delta}_{IPW2}$ and $\hat{\Delta}_{IPW3}$, which is reflected in dramatically improved efficiencies relative to $\hat{\Delta}_S$. In scenarios involving strong association between $\mathbf{X}$ and treatment exposure, $\hat{\Delta}_{SR}$ outperforms $\hat{\Delta}_{DR}$, with smaller variance and hence higher relative efficiency; otherwise, these two estimators exhibit approximately equivalent performance. Consistent with its 'benchmark' role, the ML

Table I. Monte Carlo results, multivariate confounder, correct regression modelling. Bias_S is bias of $\hat{\Delta}_S$ (per cent of true value $\Delta_0 = 2.0$). For each $(\boldsymbol{\xi}, \boldsymbol{\beta})$ setting, (i) denotes estimators using $\mathbf{X}$ only, (ii) denotes estimators using $\mathbf{X}$ and $\mathbf{V}$ as in Section 3. MC MSE ratios are computed as MC MSE_S/MC MSE_m, where m denotes the indicated estimator and MC MSE is MC bias squared plus MC variance.

				MC standard deviation						MSE ratio				
ξ	β		Bias_S	$\hat{\Delta}_S$	$\hat{\Delta}_{SR}$	$\hat{\Delta}_{IPW2}$	$\hat{\Delta}_{IPW3}$	$\hat{\Delta}_{DR}$	$\hat{\Delta}_{ML}$	*SR*	*IPW*2	*IPW*3	*DR*	*ML*
$n = 1000$														
ξ^{str}	β^{str}	(i)	−28.4	0.184	0.151	0.454	0.234	0.167	0.134	15.65	1.73	5.92	12.79	19.91
		(ii)	−28.5	0.151	0.087	0.450	0.208	0.097	0.077	45.80	1.72	7.01	37.03	59.38
	β^{mod}	(i)	−16.0	0.153	0.118	0.150	0.138	0.119	0.117	8.99	5.59	6.61	8.85	9.28
		(ii)	−15.9	0.125	0.072	0.120	0.103	0.071	0.069	22.65	8.09	11.01	22.83	24.47
ξ^{mod}	β^{str}	(i)	−22.3	0.136	0.106	0.356	0.180	0.116	0.093	19.41	1.71	5.92	16.25	25.05
		(ii)	−22.6	0.128	0.089	0.361	0.175	0.099	0.078	27.81	1.68	6.25	22.39	36.11
	β^{mod}	(i)	−12.7	0.111	0.083	0.112	0.100	0.083	0.082	11.26	6.19	7.66	11.12	11.46
		(ii)	−12.8	0.103	0.070	0.103	0.089	0.070	0.068	15.32	7.17	9.44	15.56	16.40
ξ^{no}	β^{str}	(i)	−16.1	0.109	0.091	0.252	0.138	0.098	0.080	13.80	1.81	5.43	11.97	17.86
		(ii)	−16.1	0.111	0.092	0.263	0.140	0.099	0.080	13.66	1.67	5.35	11.89	17.96
	β^{mod}	(i)	− 9.0	0.088	0.069	0.090	0.081	0.069	0.067	8.35	4.96	6.04	8.32	8.71
		(ii)	− 9.0	0.086	0.069	0.091	0.082	0.069	0.067	8.27	4.83	5.93	8.24	8.68
$n = 5000$														
ξ^{str}	β^{str}	(i)	−28.5	0.079	0.064	0.206	0.110	0.070	0.059	80.22	7.75	26.1	67.0	95.15
		(ii)	−28.5	0.067	0.039	0.203	0.102	0.042	0.035	219.05	8.00	30.10	183.10	265.80
	β^{mod}	(i)	−16.2	0.067	0.052	0.066	0.061	0.052	0.051	40.93	25.40	29.50	40.60	41.73
		(ii)	−16.1	0.051	0.030	0.050	0.044	0.030	0.030	118.81	42.20	55.00	119.20	121.57
ξ^{mod}	β^{str}	(i)	−22.3	0.061	0.047	0.168	0.088	0.052	0.043	92.57	7.16	25.30	73.70	112.09
		(ii)	−22.4	0.057	0.039	0.168	0.084	0.045	0.035	130.78	7.23	27.20	102.00	162.67
	β^{mod}	(i)	−12.6	0.052	0.038	0.050	0.046	0.039	0.038	44.79	26.90	31.80	43.70	45.28
		(ii)	−12.7	0.046	0.031	0.043	0.039	0.031	0.031	70.00	35.10	43.70	68.90	70.89
ξ^{no}	β^{str}	(i)	−16.1	0.047	0.038	0.118	0.065	0.042	0.034	73.03	7.52	24.00	60.70	92.28
		(ii)	−16.1	0.048	0.038	0.119	0.065	0.042	0.034	73.25	7.49	24.10	60.80	92.32
	β^{mod}	(i)	− 9.2	0.039	0.031	0.038	0.036	0.031	0.031	36.75	24.40	27.80	36.50	37.82
		(ii)	− 9.2	0.040	0.031	0.038	0.036	0.031	0.031	36.61	24.10	27.50	36.30	37.67

estimator exceeds (under $\boldsymbol{\beta}^{str}$) or attains similar performance to (under $\boldsymbol{\beta}^{mod}$) that of $\hat{\Delta}_{DR}$ and $\hat{\Delta}_{SR}$.

Comparison of results under (i) and (ii) confirm the reduction in variance expected from the theory in Section 3.3 for weighted estimators when 'over-fitting' the propensity score using prognostic covariates, i.e. when $\boldsymbol{\xi} = \boldsymbol{\xi}^{mod}$ or $\boldsymbol{\xi}^{str}$. The few instances of slight efficiency loss

at $n=1000$ are resolved at $n=5000$. Gains achieved by $\hat{\Delta}_{DR}$ are most dramatic. Moreover, for a particular $\boldsymbol{\beta}$ setting, including $\mathbf{V}$ in the analysis with $\hat{\Delta}_{DR}$ when $\boldsymbol{\xi}=\boldsymbol{\xi}^{\text{mod}}$ or $\boldsymbol{\xi}^{\text{str}}$ results in MC standard deviation equal to that possible when there is no association between $\mathbf{V}$ and response ($\boldsymbol{\xi}=\boldsymbol{\xi}^{\text{no}}$). In contrast, the other weighted estimators gain efficiency by including $\mathbf{V}$, but an increase in the magnitude of $\boldsymbol{\xi}$ is associated with an increase in variance. Although theory in Section 3.3 is not informative for $\hat{\Delta}_S$ and $\hat{\Delta}_{SR}$, the empirical results suggest that their sampling variation is also reduced by such 'over-fitting'. In fact, we have evaluated $\boldsymbol{\Gamma}_{\gamma\beta qp}$ in (34) in numerous situations and found its sign always to be negative.

Table II shows analogous results for $\hat{\Delta}_{SR^*}, \hat{\Delta}_{DR^*}$, and $\hat{\Delta}_{ML^*}$. 'Double robustness' of $\hat{\Delta}_{DR^*}$ is confirmed; under all scenarios, the bias of this estimator is less than 1 per cent and is thus not shown. Moreover, the efficiency of this estimator relative to $\hat{\Delta}_{DR}$, which uses correct regression models, only suffers noticeably when $\boldsymbol{\beta}=\boldsymbol{\beta}^{\text{str}}$ and is superior to that of $\hat{\Delta}_{IPW2}$ and $\hat{\Delta}_{IPW3}$ in every case, showing that 'augmentation' of usual weighted estimators by regression relationships may increase precision even if the models are not exactly correct. In contrast, failure to incorporate the correct regression relationship leads to bias of $\hat{\Delta}_{SR^*}$, although its magnitude is smaller than that of $\hat{\Delta}_S$ in Table I. This feature results in considerably poorer efficiency of $\hat{\Delta}_{SR^*}$ relative to $\hat{\Delta}_{DR^*}$. The drawback of direct regression modelling is clearly evident; using an incorrect model yields significant bias and consequently drastically inferior performance. These results suggest that, if one insists on estimators like $\hat{\Delta}_{SR}$ or $\hat{\Delta}_{ML}$ that involve regression modelling explicitly, the former is 'safer.' The nature of the mismodelling we have examined was chosen deliberately to be rather extreme to demonstrate the potential pitfalls of these approaches; here, disregarding X_1 in the regression modelling disregards a confounder, emphasizing how sensitive these estimators are to violation of key assumptions in the regression model, a situation to which $\hat{\Delta}_{DR}$ is robust.

To further assess the quality of inference, we calculated nominal 95 per cent Wald confidence intervals for Δ_0 as estimate $\pm 1.96\times$ estimated standard deviation for each estimator, using the sandwich method based on (18)–(21) for the weighted estimators, using (29) for $\hat{\Delta}_S$ and the analogous approach for $\hat{\Delta}_{SR}$, and using the usual *OLS* standard error for $\hat{\Delta}_{ML}$. Table III shows Monte Carlo coverage probabilities for case (i). Low coverages for $\hat{\Delta}_S$ are due to the residual biases in Table I, as estimated standard errors from (29) performed well, closely tracking the MC standard deviations. Coverage for $\hat{\Delta}_{IPW2}$ and $\hat{\Delta}_{IPW3}$ achieves the nominal level under $\boldsymbol{\beta}^{\text{mod}}$, with somewhat optimistic performance when this association is strong. Notably, coverages for $\hat{\Delta}_{DR}, \hat{\Delta}_{SR}$, and $\hat{\Delta}_{ML}$ attain the nominal level in all cases; moreover, so do those for $\hat{\Delta}_{DR^*}$, despite augmentation by the 'wrong' regression model. In contrast, due to the biases in Table II, coverages based on $\hat{\Delta}_{SR^*}$ and $\hat{\Delta}_{ML^*}$ are far from nominal.

The foregoing results take $K=5$ for $\hat{\Delta}_S$, as is common in practice; however, with larger sample sizes, one might refine the balancing effect of stratification by increasing K. Table IV shows for case (i) performance of $\hat{\Delta}_S$ when the number of strata was doubled from $K=5$ to 10. While MC standard deviations and standard errors for $\hat{\Delta}_S$ are similar and remain fairly constant from $K=5$ to 10, bias is reduced by roughly 65 per cent in all scenarios, yielding improved coverage (although still not at the nominal level). However, performance of $\hat{\Delta}_S$ is still inferior to that of the other estimators, and, because residual bias, although smaller than for $K=5$, remains constant as n increases, coverage worsens for $n=5000$.

Table II. Monte Carlo results, multivariate confounder, incorrect regression modelling. Bias_{SR^*} and Bias_{ML^*} are bias of $\hat{\Delta}_{SR^*}$ and $\hat{\Delta}_{ML^*}$ (percent of true value $\Delta_0 = 2.0$). All other entries are as in Table I.

					MC standard deviation			MSE ratio		
ξ	β		Bias_{SR^*}	Bias_{ML^*}	$\hat{\Delta}_{SR^*}$	$\hat{\Delta}_{DR^*}$	$\hat{\Delta}_{ML^*}$	SR^*	DR^*	ML^*
$n = 1000$										
ξ^{str}	β^{str}	(i)	−11.9	−35.2	0.166	0.207	0.164	4.24	8.30	0.68
		(ii)	− 8.3	−23.6	0.107	0.141	0.120	8.96	17.53	1.47
	β^{mod}	(i)	− 6.7	−18.0	0.131	0.121	0.152	3.62	8.55	0.83
		(ii)	− 4.5	−12.0	0.085	0.074	0.109	7.69	21.52	1.67
ξ^{mod}	β^{str}	(i)	− 9.8	−28.4	0.118	0.141	0.124	4.17	10.99	0.64
		(ii)	− 7.8	−21.5	0.102	0.121	0.106	6.36	15.03	1.12
	β^{mod}	(i)	− 5.3	−14.7	0.092	0.085	0.110	3.89	10.57	0.78
		(ii)	− 4.2	−11.2	0.077	0.072	0.094	5.88	14.79	1.28
ξ^{no}	β^{str}	(i)	− 7.3	−21.0	0.103	0.118	0.101	3.61	8.34	0.62
		(ii)	− 6.8	−18.8	0.101	0.118	0.100	4.05	8.40	0.77
	β^{mod}	(i)	− 3.8	−10.9	0.075	0.070	0.087	3.58	8.08	0.73
		(ii)	− 3.5	− 9.6	0.073	0.070	0.085	3.94	8.03	0.90
$n = 5000$										
ξ^{str}	β^{str}	(i)	−12.2	−35.3	0.069	0.084	0.074	5.15	46.74	0.65
		(ii)	− 8.6	−23.7	0.047	0.058	0.055	10.32	98.13	1.45
	β^{mod}	(i)	− 6.9	−18.3	0.056	0.053	0.065	4.93	39.50	0.79
		(ii)	− 4.8	−12.2	0.035	0.031	0.049	10.32	114.02	1.72
ξ^{mod}	β^{str}	(i)	− 9.9	−28.4	0.052	0.067	0.058	4.83	44.57	0.63
		(ii)	− 7.9	−21.4	0.045	0.056	0.049	7.62	64.86	1.10
	β^{mod}	(i)	− 5.5	−14.7	0.042	0.039	0.050	4.83	42.89	0.74
		(ii)	− 4.3	−11.1	0.034	0.031	0.043	7.82	68.23	1.30
ξ^{no}	β^{str}	(i)	− 7.4	−21.3	0.041	0.053	0.044	4.50	37.79	0.58
		(ii)	− 6.9	−19.1	0.042	0.052	0.043	5.06	39.55	0.72
	β^{mod}	(i)	− 4.1	−11.1	0.034	0.032	0.040	4.46	35.36	0.71
		(ii)	− 3.8	− 9.9	0.034	0.032	0.039	5.09	35.19	0.88

5. DISCUSSION

We have reviewed and compared two principal approaches to estimating average causal effects from observational data using the propensity score, those based on stratification and weighting. We hope that this presentation serves as a resource to practitioners who wish to appreciate the rationale for and differences between these two classes of techniques and to understand

Table III. Monte Carlo coverage probabilities for case (i) in Tables I and II.

ξ	β	$\hat{\Delta}_S$	$\hat{\Delta}_{SR}$	$\hat{\Delta}_{SR^*}$	$\hat{\Delta}_{IPW2}$	$\hat{\Delta}_{IPW3}$	$\hat{\Delta}_{DR}$	$\hat{\Delta}_{DR^*}$	$\hat{\Delta}_{ML}$	$\hat{\Delta}_{ML^*}$
$n=1000$										
ξ^{str}	β^{str}	13.5	94.7	71.5	88.4	88.0	94.5	94.3	94.6	1.3
	β^{mod}	44.8	94.8	83.6	94.1	93.6	94.9	95.1	94.6	32.8
ξ^{mod}	β^{str}	9.8	95.4	68.1	88.1	87.3	95.8	94.5	95.2	0.2
	β^{mod}	38.1	95.0	82.6	94.9	93.9	95.3	95.1	95.0	26.1
ξ^{no}	β^{str}	15.1	94.1	70.6	89.2	88.9	94.8	93.9	95.3	1.7
	β^{mod}	49.1	95.6	85.4	94.6	94.7	95.7	95.5	95.6	32.5
$n=5000$										
ξ^{str}	β^{str}	0.0	95.3	9.0	91.5	91.5	95.6	95.2	94.7	0.0
	β^{mod}	0.1	95.7	37.5	95.6	95.2	95.9	95.7	95.8	0.0
ξ^{mod}	β^{str}	0.0	94.9	4.6	91.0	90.8	94.3	93.2	95.0	0.0
	β^{mod}	0.1	94.3	28.8	94.9	94.5	94.5	95.0	94.0	0.0
ξ^{no}	β^{str}	0.0	95.4	8.6	91.5	90.3	95.6	93.9	96.4	0.0
	β^{mod}	0.3	95.1	34.8	95.5	95.7	94.9	94.4	94.8	0.0

Table IV. Monte Carlo results for $\hat{\Delta}_S$ at $K=10$ for case (i) Table I. Bias is bias of $\hat{\Delta}_S$ expressed as percentage of the true value $\Delta_0=2.0$. MC SD is Monte Carlo standard deviation, Ave SE is the average of estimated standard errors of $\hat{\Delta}_S$ using (29), and Coverage is Monte Carlo coverage of 95 per cent confidence interval. MSE ratios are as in Table I; $\hat{\Delta}_{SR}$ is still based on $K=5$ as in previous tables.

						MSE ratio			
ξ	β	Bias	MC SD	(Ave SE)	Coverage	*IPW2*	*IPW3*	*DR*	*SR*
$n=1000$									
ξ^{str}	β^{str}	−9.9	0.188	(0.167)	72.9	0.39	1.23	2.82	3.28
	β^{mod}	−5.3	0.133	(0.135)	88.4	1.26	1.50	2.05	2.06
ξ^{mod}	β^{str}	−7.9	0.141	(0.122)	72.4	0.34	1.32	3.55	3.98
	β^{mod}	−4.4	0.099	(0.098)	85.0	1.39	1.69	2.40	2.58
ξ^{no}	β^{str}	−6.0	0.111	(0.097)	73.9	0.39	1.25	2.95	3.18
	β^{mod}	−3.2	0.077	(0.078)	87.7	1.35	1.56	2.09	3.09
$n=5000$									
ξ^{str}	β^{str}	−10.0	0.077	(0.076)	25.0	0.99	3.38	8.78	11.21
	β^{mod}	−5.5	0.059	(0.059)	53.1	3.57	4.13	5.62	5.92
ξ^{mod}	β^{str}	−7.7	0.055	(0.055)	19.3	1.07	3.82	10.04	12.17
	β^{mod}	−4.3	0.042	(0.043)	48.1	3.84	4.55	6.58	6.20
ξ^{no}	β^{str}	−5.7	0.047	(0.045)	26.9	1.14	3.40	7.97	10.64
	β^{mod}	−3.1	0.035	(0.034)	54.8	3.29	3.81	5.20	5.27

their relative performance. Strategies based on matching on propensity scores or adjusting for the propensity score in direct regression modelling [2], which we did not consider, are also popular.

Theoretical and empirical results indicate that the popular version of stratification via estimated propensity scores based on within-stratum sample mean differences and a fixed number

of strata can lead to biased inference due to residual confounding, and the effect of this bias becomes more serious with increasing sample size. Using more strata can increase the sample size at which the trade-off of bias and variability involved in efficiency takes place, but stratifying on quintiles seems to be the most popular approach in practice, even for substantial sample sizes. Thus, as the 'trade-off' point will be unknown for any specific problem, this approach should be used with caution. An interesting avenue for future research would be to establish guidelines for choosing the number of strata based on theoretical analysis of the rate at which the number of strata should increase with sample size to eliminate bias. A modification of stratification based instead on within-stratum regression estimates of treatment effect can eliminate this bias and achieve dramatic improvements in efficiency, but correct specification of the regression model is essential; otherwise, bias and degradation of performance can result. In this regard, this approach is similar to estimating causal effects via direct regression modelling but is less sensitive to mismodelling.

Methods based on weighting are consistent and offer approximately unbiased inference for practical sample sizes. The semiparametric efficient estimator identified by the theory of Robins *et al.* [13], which incorporates regression modelling as a way to gain efficiency, also yields high precision. Although stratification based on regression and direct modelling can outperform this approach under some conditions, this estimator enjoys the unique 'double robustness' property in that it continues to lead to unbiased estimation of the average causal effect even if the regression models involved do not coincide with the true relationship, affording the analyst broad protection against misspecification not available with these other approaches. The results presented here support routine use of this estimator in practice.

APPENDIX A: DERIVATION OF (27) AND (34)

Applying the results in Section A.3.6 of Reference [22] to (26), we have

$$\sum_S = A_{22}^{-1}(\mathbf{B}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{B}_{12} - \mathbf{B}_{12}^{\mathrm{T}}\mathbf{A}_{11}^{-\mathrm{T}}\mathbf{A}_{21}^{\mathrm{T}} + \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{B}_{11}\mathbf{A}_{11}^{-\mathrm{T}}\mathbf{A}_{21}^{\mathrm{T}})\mathbf{A}_{22}^{-\mathrm{T}} \tag{A36}$$

where the matrices in this expression follow from tedious evaluation of the required derivatives and covariance matrix. In particular, it may be shown that $A_{22} = -1$, and

$$\mathbf{A}_{11} = \begin{pmatrix} \mathbf{E}_{qq} & \mathbf{0} & \mathbf{E}_{q\beta} \\ \mathbf{E}_{pq} & -\mathbf{I}_K & \mathbf{E}_{p\beta} \\ \mathbf{0} & \mathbf{0} & -\mathbf{E}_{\beta\beta} \end{pmatrix}, \quad \mathbf{B}_{11} = \begin{pmatrix} \mathbf{F}_{qq} & \mathbf{F}_{qp} & \mathbf{0} \\ \mathbf{F}_{qp}^{\mathrm{T}} & \mathbf{F}_{pp} & \mathbf{F}_{p\beta} \\ \mathbf{0} & \mathbf{F}_{p\beta}^{\mathrm{T}} & \mathbf{E}_{\beta\beta} \end{pmatrix}$$

Here, $\mathbf{E}_{qq} = \mathrm{diag}\{f_e(q_1), f_e(q_2), \ldots, f_e(q_{K-1})\}$; $\mathbf{E}_{pq}^{(i,j)} = q_j f_e(q_j)$, $i = j, -q_j f_e(q_j), i = j+1$, and zero otherwise $(K \times K-1)$; and $\mathbf{E}_{q\beta}(K-1 \times p)$ has jth row $\partial/\partial\boldsymbol{\beta}^{\mathrm{T}}\{\int_0^{q_j} f_e(t)\,\mathrm{d}t\}$ and $\mathbf{E}_{p\beta}(K \times p)$ has jth row $\partial/\partial\boldsymbol{\beta}^{\mathrm{T}}\{\int_{q_{j-1}}^{q_j} t f_e(t)\,\mathrm{d}t\}$, where differentiation is with respect to $\boldsymbol{\beta}$ in $f_e(\cdot)$ only. In addition, $\mathbf{F}_{qq}$ is symmetric with (i,j) upper-triangular element $(i/K)(1-j/K)$; $\mathbf{F}_{qp}^{(i,j)} = p_j(1-i/K)$, $i \geqslant j, = -p_j(i/K), i < j(K-1 \times K)$; $\mathbf{F}_{pp}(K \times K)$ is symmetric with $\mathbf{F}_{pp}^{(j,j)} = p_j(1-p_j), \mathbf{F}_{pp}^{(i,j)} = -p_i p_j$; and $\mathbf{F}_{p\beta}(K \times p)$ has jth row $E\{I(e \in Q_j)e_\beta^{\mathrm{T}}\}$, where the expectation is with respect to the distribution of $\mathbf{X}$. Defining $h_{1j} = p_j^{-1}\int_{q_{j-1}}^{q_j} E(Y_1 \mid t) t f_e(t)\,\mathrm{d}t$ and $h_{0j} = (1/K - p_j)^{-1}\int_{q_{j-1}}^{q_j} E(Y_0 \mid t)$

$(1-t)f_e(t)\,\mathrm{d}t$, $j=1,\ldots,K$, and $g_{1j}=E(Y_1\,|\,q_j)q_j(p_j^{-1}-p_{j+1}^{-1})$ and $g_{0j}=E(Y_0\,|\,q_j)(1-q_j)\{(1/K-p_j)^{-1}-(1/K-p_{j+1})^{-1}\}$, $j=1,\ldots,K-1$, then $\mathbf{A}_{21}=(\mathbf{E}_{\Delta q}\ \mathbf{E}_{\Delta p}\ \mathbf{E}_{\Delta\beta})$, $\mathbf{B}_{12}^{\mathrm{T}}=(\mathbf{F}_{q\Delta}^{\mathrm{T}}\ \mathbf{F}_{p\Delta}^{\mathrm{T}}\ \mathbf{F}_{\beta\Delta}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{E}_{\Delta p}(1\times K)$ has jth element $(p_jK)^{-1}h_{1j}-(1-K_{p_j})^{-1}h_{0j}$, respectively; $\mathbf{E}_{\Delta q}(1\times K-1)$ has elements $K^{-1}(g_{1j}-g_{0j})f_e(q_j)$; and $\mathbf{E}_{\Delta\beta}(1\times p)$ is given by

$$\partial/\partial\boldsymbol{\beta}^{\mathrm{T}}\left[\sum_{j=1}^{K}\left\{(p_jK)^{-1}\int_{q_{j-1}}^{q_j}E(Y_1\,|\,t)tf_e(t)\,\mathrm{d}t-(K^{-1}-p_j)^{-1}\int_{q_{j-1}}^{q_j}E(Y_0\,|\,t)(1-t)f_e(t)\,\mathrm{d}t\right\}\right]$$

where differentiation is with respect to $\boldsymbol{\beta}$ in $f_e(\cdot)$. Similarly, $\mathbf{F}_{p\Delta}^{\mathrm{T}}(1\times K)$ has jth element $K^{-1}h_{1j}-p_j\Delta^*$; $\mathbf{F}_{q\Delta}^{\mathrm{T}}(1\times K-1)$ has elements $K^{-1}\sum_{i=1}^{j}(h_{1i}-h_{0i}-\Delta^*)$; and $\mathbf{F}_{\beta\Delta}^{\mathrm{T}}(1\times p)$ is $K^{-1}\sum_{j=1}^{K}[p_j^{-1}E\{Y_1I(e\in Q_j)e_\beta^{\mathrm{T}}\}+(1/K-p_j)^{-1}E\{Y_0I(e\in Q_j)e_\beta^{\mathrm{T}}\}]$.

Substituting these expressions in (36) and simplifying yields (27), with $\boldsymbol{\Gamma}_p=\mathbf{E}_{\Delta p}\mathbf{F}_{p\Delta}+\mathbf{F}_{p\Delta}^{\mathrm{T}}\mathbf{E}_{\Delta p}^{\mathrm{T}}+\mathbf{E}_{\Delta p}\mathbf{F}_{pp}\mathbf{E}_{\Delta p}^{\mathrm{T}}$, $\boldsymbol{\Gamma}_{qp}=-\mathbf{H}_{\Delta q}(\mathbf{E}_{\Delta p}\mathbf{F}_{qp}^{\mathrm{T}}+\mathbf{F}_{q\Delta}^{\mathrm{T}})^{\mathrm{T}}-(\mathbf{E}_{\Delta p}\mathbf{F}_{qp}^{\mathrm{T}}+\mathbf{F}_{q\Delta}^{\mathrm{T}})\mathbf{H}_{\Delta q}^{\mathrm{T}}+\mathbf{H}_{\Delta q}\mathbf{F}_{qq}\mathbf{H}_{\Delta q}^{\mathrm{T}}$, and $\boldsymbol{\Gamma}_{\beta qp}=(\mathbf{H}_{\Delta\beta}-\mathbf{H}_{\Delta q}\mathbf{E}_{q\beta})\mathbf{E}_{\beta\beta}^{-1}(F_{\beta\Delta}^{\mathrm{T}}+\mathbf{E}_{\Delta p}\mathbf{F}_{\beta p}^{\mathrm{T}})^{\mathrm{T}}+(F_{\beta\Delta}^{\mathrm{T}}+\mathbf{E}_{\Delta p}\mathbf{F}_{\beta p}^{\mathrm{T}})\mathbf{E}_{\beta\beta}^{-1}(\mathbf{H}_{\Delta\beta}-\mathbf{H}_{\Delta q}\mathbf{E}_{q\beta})^{\mathrm{T}}+(\mathbf{H}_{\Delta\beta}-\mathbf{H}_{\Delta q}\mathbf{E}_{q\beta})\mathbf{E}_{\beta\beta}^{-1}(\mathbf{H}_{\Delta\beta}-\mathbf{H}_{\Delta q}\mathbf{E}_{q\beta})^{\mathrm{T}}$, where $\mathbf{H}_{\Delta q}=(\mathbf{E}_{\Delta q}+\mathbf{E}_{\Delta p}\mathbf{E}_{pq})\mathbf{E}_{qq}^{-1}$ and $\mathbf{H}_{\Delta\beta}=\mathbf{E}_{\Delta\beta}+\mathbf{E}_{\Delta p}\mathbf{E}_{p\beta}$.

To obtain the second term in (34), let $\mathbf{E}_{\Delta\gamma}(1\times q)$ equal

$$\partial/\partial\boldsymbol{\gamma}^{\mathrm{T}}\left[\sum_{j=1}^{K}\left\{(p_jK)^{-1}\int_{q_{j-1}}^{q_j}E(Y_1\,|\,t)tf_e(t)\,\mathrm{d}t-(K^{-1}-p_j)^{-1}\int_{q_{j-1}}^{q_j}E(Y_0\,|\,t)(1-t)f_e(t)\,\mathrm{d}t\right\}\right]$$

Let $\mathbf{E}_{q\gamma}(K-1\times q)$ and $\mathbf{E}_{p\gamma}(K\times q)$ have jth rows $\partial/\partial\boldsymbol{\gamma}^{\mathrm{T}}\left\{\int_0^{q_j}f_e(t)\,\mathrm{d}t\right\}$ and $\partial/\partial\boldsymbol{\gamma}^{\mathrm{T}}\left\{\int_{q_{j-1}}^{q_j}tf_e(t)\,\mathrm{d}t\right\}$, respectively. Also let $\mathbf{F}_{p\gamma}(K\times q)$ be the matrix with jth row $E\{I(e\in Q_j)e_\gamma^{\mathrm{T}}\}$, and $\mathbf{F}_{\gamma\Delta}^{\mathrm{T}}(1\times p)$ is $K^{-1}\sum_{j=1}^{K}[p_j^{-1}E\{Y_1I(e\in Q_j)e_\gamma^{\mathrm{T}}\}+(1/K-p_j)^{-1}E\{Y_0I(e\in Q_j)e_\gamma^{\mathrm{T}}\}]$. Defining $\mathbf{H}_{\Delta\gamma}=\mathbf{E}_{\Delta\gamma}-\mathbf{E}_{\Delta p}\mathbf{E}_{p\gamma}$, $\mathbf{D}_\gamma=\mathbf{H}_{\Delta\gamma}-\mathbf{H}_{\Delta\beta}\mathbf{E}_{\beta\beta}^{-1}\mathbf{E}_{\gamma\beta}^{\mathrm{T}}-\mathbf{H}_{\Delta q}(\mathbf{E}_{q\gamma}-\mathbf{E}_{q\beta}\mathbf{E}_{\beta\beta}^{-1}\mathbf{E}_{\gamma\beta}^{\mathrm{T}})$, and $\mathbf{G}_\gamma=(\mathbf{F}_{\gamma\Delta}-\mathbf{E}_{\gamma\beta}\mathbf{E}_{\beta\beta}^{-1}\mathbf{F}_{\beta\Delta})^{\mathrm{T}}+\mathbf{E}_{\Delta p}(\mathbf{F}_{p\gamma}^{\mathrm{T}}-\mathbf{E}_{\gamma\beta}\mathbf{E}_{\beta\beta}^{-1}\mathbf{E}_{\beta p})^{\mathrm{T}}$, one can show that $\Gamma_{\gamma\beta qp}=\mathbf{D}_\gamma\mathbf{H}_{\gamma\beta}^{-1}\mathbf{G}_\gamma^{\mathrm{T}}+\mathbf{G}_\gamma\mathbf{H}_{\gamma\beta}^{-1}\mathbf{D}_\gamma^{\mathrm{T}}+\mathbf{D}_\gamma\mathbf{H}_{\gamma\beta}^{-1}\mathbf{D}_\gamma^{\mathrm{T}}$.

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. Tutorial in biostatistics: 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. Propensity score. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds), vol. 5. Wiley: New York, 1998; 3551–3555.
4. Shepardson LB, Youngner SJ, Speroff T, Rosenthal GE. Increased risk of death in patients with do-not-resuscitate orders. *Medical Care* 1999; **37**:727–737.
5. Perkins SM, Tu W, Underhill MG, Zhou XH, Murray MD. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety* 2000; **9**:93–101.
6. Allen-Ramey FC, Duong PT, Goodman DC, Sajjan SG, Nelsen LM, Santanello NC, Markson LE. Treatment effectiveness of inhaled corticosteroids and leukotriene modifiers for patients with asthma: an analysis from managed care data. *Allergy and Asthma Proceedings* 2003; **24**:43–51.
7. 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.
8. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987; **82**: 387–394.
9. Robins JM, Hernán M, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**:550–560.
10. Rubin DR. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.

11. Polsky D, Mandelblatt JS, Weeks JC, Venditti L, Hwang Y, Glick HA, Hadley J, Schulman KA. Economic evaluation of breast cancer treatment: considering the value of patient choice. *Journal of Clinical Oncology* 2003; **21**:1139–1146.
12. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**:663–685.
13. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846–866.
14. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* 1999; 6–10.
15. Scharfstein DO, Rotnitzky A, Robins JM. Rejoinder to Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* 1999; **94**:1135–1146.
16. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* 2001; **2**:259–278.
17. Rubin DR. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997; **127**:757–763.
18. Stefanski LA, Boos DD. The calculus of M-estimation. *The American Statistician* 2002; **56**:29–38.
19. McIntosh MW, Rubin DB. On estimating the causal effects of DNR orders. *Medical Care* 1999; **37**:722–726.
20. Dawid AP. Conditional independence in statistical theory. *Journal of the Royal Statistical Society*, *Series B* 1979; **41**:1–31.
21. Lunceford JK. Estimating causal treatment effects via the propensity score and estimating survival distributions in clinical trials that follow two-stage randomization designs. Unpublished *Ph.D. dissertation*, North Carolina State University, 2001; available at http://www.lib.ncsu.edu/.
22. Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models*. Chapman & Hall, London; 1995.

