Causal inference with a quantitative exposure

Article *in* Statistical Methods in Medical Research · June 2012

DOI: 10.1177/0962280212452333 · Source: PubMed

4 authors, including:

Zhiwei Zhang
National Cancer Institute
101 PUBLICATIONS **1,241** CITATIONS

Statistical Methods in Medical Research
http://smm.sagepub.com/

Causal inference with a quantitative exposure

Zhiwei Zhang, Jie Zhou, Weihua Cao and Jun Zhang

Stat Methods Med Res published online 22 June 2012
DOI: 10.1177/0962280212452333

The online version of this article can be found at:
http://smm.sagepub.com/content/early/2012/06/22/0962280212452333

Published by:
SAGE
http://www.sagepublications.com

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–21

DOI: 10.1177/0962280212452333
smm.sagepub.com
SAGE

Causal inference with a quantitative exposure

Zhiwei Zhang,[1] Jie Zhou,[1] Weihua Cao[1] and Jun Zhang[2]

Abstract

The current statistical literature on causal inference is mostly concerned with binary or categorical exposures, even though exposures of a quantitative nature are frequently encountered in epidemiologic research. In this article, we review the available methods for estimating the dose–response curve for a quantitative exposure, which include ordinary regression based on an outcome regression model, inverse propensity weighting and stratification based on a propensity function model, and an augmented inverse propensity weighting method that is doubly robust with respect to the two models. We note that an outcome regression model often imposes an implicit constraint on the dose–response curve, and propose a flexible modeling strategy that avoids constraining the dose–response curve. We also propose two new methods: a weighted regression method that combines ordinary regression with inverse propensity weighting and a stratified regression method that combines ordinary regression with stratification. The proposed methods are similar to the augmented inverse propensity weighting method in the sense of double robustness, but easier to implement and more generally applicable. The methods are illustrated with an obstetric example and compared in simulation studies.

Keywords

Dose–response relationship, double robustness, inverse probability weighting, outcome regression, propensity function, propensity score, stratification

1 Introduction

Understanding the causal effect of a treatment or exposure is an important objective in many epidemiologic studies. A common framework for understanding causality is Rubin's[1] causal model, and a standard assumption is strongly ignorable treatment assignment.[2] Under this assumption, valid causal inference can be based on an outcome regression model that relates the

[1]Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland, USA
[2]MOE and Shanghai Key Laboratory of Children's Environmental Health, Xinhua Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, P.R. China

Corresponding author:
Zhiwei Zhang, Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland, USA.
Email: zhiwei_zhang@yahoo.com; zhiwei.zhang@fda.hhs.gov

outcome of interest to the treatment or exposure under investigation as well as observed confounders. This ordinary regression (OR) approach is straightforward to implement and the resulting inference is efficient (under a correctly specified parametric model). Many alternative methods require a model for the propensity score, that is, the conditional probability of being treated or exposed given covariate values.[2] Estimated propensity scores can be used for matching, stratification, or inverse probability weighting.[3–6] Much of the recent research has been directed toward doubly robust estimators that involve both an outcome regression model and a propensity score model and that remain consistent if one of the models (but not both) is misspecified.[6–18]

Most of the existing literature on causal inference is concerned with binary or categorical exposures. In practice, exposures of a quantitative nature are frequently encountered. For example, although smoking is often treated as a binary exposure, information is often available on the amount of smoking (e.g. frequency times duration) for a smoker. As another example, obesity is usually defined by dichotomizing the body mass index (BMI), a continuous variable that is arguably more informative than a dichotomous indicator of obesity. This article gives a overview of the existing methods for estimating a does–response curve that summarizes the effect of a quantitative exposure. The OR approach mentioned earlier can accommodate any type of exposures. We note, however, that an outcome regression model can impose an implicit constraint on the dose–response curve, and propose a flexible modeling strategy that does not constrain the dose–response curve. The propensity score, originally defined for a binary exposure, can be generalized into a propensity function for a quantitative exposure.[19,20] Robins et al.[19] show that a marginal structural model (MSM) for a quantitative exposure can be estimated consistently using an inverse propensity weighting (IPW) method, weighting each subject by the inverse of the estimated propensity function. Following Joffe and Rosenbaum,[21] Imai and Van Dyk[20] assume that the propensity function has a certain structure that allows dimension reduction and propose a stratification method that is approximately consistent if the propensity function model is correct and the strata are approximately homogeneous. Each of the aforementioned methods for a quantitative exposure is based on either an outcome regression model or a propensity function model. These models are typically parametric or at least semiparametric due to the curse of dimensionality, raising concerns about potential bias due to model misspecification.

For a parametric MSM, Van Der Laan and Robins[10] develop an augmented IPW (AIPW) method that is doubly robust and locally efficient. The double robustness property of this method appears limited to special models by a compatibility requirement for the MSM and the outcome regression model. Furthermore, it can be challenging to implement the AIPW method, which involves an integral in the estimating function. In this article, we propose two methods that are similar to the AIPW method in the sense of double robustness, but are easier to implement and more generally applicable. Specifically, we propose a weighted regression (WR) method that combines OR with IPW and a stratified regression method that combines OR with stratification. For estimating a suitably parameterized or completely non-parametric MSM, the WR method is doubly robust and the stratified regression method is approximately so (consistent under the outcome regression model and approximately consistent under the propensity function model). Although the stratified regression method is theoretically somewhat inferior to the AIPW and WR methods, it is more protected from very large weights, a common problem with IPW.

The rest of the article is organized as follows. In the next section, we set up the notation and state the key assumptions. We then give an overview of the existing methods in Section 3, and present the proposed methods in Section 4. In Section 5, the methods are illustrated with an obstetric example and compared in simulation experiments. The article ends with a discussion in Section 6. Technical details are provided in Appendix.

2 Notation and assumptions

Let $\mathcal{T}$ denote the set of all possible values of the treatment or exposure of interest. For a binary exposure, it is customary to take $\mathcal{T} = \{0, 1\}$. For a qualitative exposure with more than two levels, one may take $\mathcal{T} = \{0, 1, \ldots, K\}$ for some $K > 1$.[21,22] Of particular interest to us is a quantitative exposure such as dose, for which $\mathcal{T}$ is usually an interval. For each $t \in \mathcal{T}$, let $Y(t)$ denote the potential outcome that would realize if a subject is exposed at level t. Suppose we are primarily interested in marginal means of the potential outcomes, then the causal effect of the exposure can be described by $\mu(t) = \mathrm{E}\{Y(t)\}$ as a function of t, often referred to as the dose–response curve.

Of course, the potential outcomes are not observed completely. In reality, we only observe the actual treatment T and the corresponding outcome $Y = Y(T)$. In randomized experiments, T is independent of the $Y(t)$, written

$$T \perp Y(t), \quad t \in \mathcal{T}.$$

This implies that

$$\mathrm{E}(Y|T = t) = \mathrm{E}\{Y(t)|T = t\} = \mathrm{E}\{Y(t)\},$$

so that $\mu(t)$ can be estimated by simply regressing Y on T, parametrically or non-parametrically. The independence of T and the potential outcomes would be difficult to justify in an observational study, where subjects at different exposure levels tend to differ systematically. Let X denote a collection of baseline variables that can be used to explain any association between the actual exposure and the potential outcomes. Formally, we assume that, for each t, $Y(t)$ is conditionally independent of T given X, that is

$$T \perp Y(t)|X, \quad t \in \mathcal{T}. \tag{1}$$

This assumption is often referred to as strong ignorability of treatment assignment or no unmeasured confounding. We also make the positivity assumption that

$$r(t|x) > 0, \quad \forall (t, x), \tag{2}$$

where $r(t|x)$ denotes the conditional density of T given $X = x$ with respect to some measure ν. It is necessary to consider a general measure ν because T may follow a mixed distribution with both continuous and discrete components. For example, smoking as a quantitative exposure has a point mass at 0 for non-smokers. The function $r(t|x)$ generalizes the propensity score for a binary exposure[2] and is known as the propensity function.[20] The essence of assumption (2) is that all exposure levels are possible for all subjects with different characteristics. Unlike assumption (1), which is not testable with the observed data, assumption (2) can and should be checked with the data, as its violations can have serious consequences.[23] Together, assumptions (1) and (2) are sufficient for non-parametric identification of $\mu(t)$ from the observable (X, T, Y). This can be seen by writing

$$\mu(t) = \mathrm{E}[\mathrm{E}\{Y(t)|X\}] = \mathrm{E}[\mathrm{E}\{Y(t)|X, T = t\}] = \mathrm{E}[\mathrm{E}\{Y|X, T = t\}], \tag{3}$$

where the middle step follows from assumption (1). Assumption (2) insures that all values of X are potentially observable with $T=t$; thus, $\mathrm{E}\{Y|X=x, T=t\}$ is identifiable for each x and so is the right-hand side of equation (3).

The observed data consist of (X_i, T_i, Y_i), $i=1,\ldots,n$, which we conceptualize as independent copies of (X, T, Y). Our objective is to estimate $\mu(t)$ using these data under assumptions (1) and (2) and appropriate modeling assumptions. We consider both parametric and non-parametric models for $\mu(t)$, recognizing that a non-parametric $\mu(t)$ cannot be estimated at the parametric rate.

3 Existing methods

3.1 Ordinary regression

The standard approach is to adjust for confounding using an outcome regression model for $\mathrm{E}(Y|T, X)$, say $m(T, X; \beta)$, where m is a known function and β an unknown parameter which is usually finite dimensional but can be infinite dimensional. Now, equation (3) can be rewritten as

$$\mu(t) = \mathrm{E}m(t, X; \beta). \tag{4}$$

Consider first the generalized linear model (GLM)

$$m(t, x; \beta) = \psi\{\beta_1 + \beta_T t + \beta'_X x + \beta'_{TX}(tx)\}, \tag{5}$$

where ψ is an inverse link function (e.g. identity for a continuous outcome, exp for count data, expit or probit for binary data). On the right-hand side, t and x may be replaced by vectors of transformations, and the interaction term is optional. For model (5), equation (4) can be further rewritten as

$$\mu(t) = \mathrm{E}\psi\{\beta_1 + \beta_T t + \beta'_X X + \beta'_{TX}(tX)\}. \tag{6}$$

This corresponds to an MSM[19] of the form

$$\mu(t; \alpha) = \psi(\alpha_1 + \alpha_T t), \tag{7}$$

where α is determined by β and the marginal distribution of X, in the following special cases:

Case A. For the identity link, expression (7) holds with $\alpha_1 = \beta_1 + \beta_X{}'\mathrm{E}\,X$ and $\alpha_T = \beta_T + \beta_{TX}{}'\mathrm{E}\,X$. For convenience, we assume in this case that X is centered (by subtracting E X), so that $\alpha = (\beta_1, \beta_T)'$.

Case B. For the log link without the interaction term, expression (7) holds with $\alpha_1 = \beta_1 + \log\{\mathrm{E}\exp(\beta_X{}'X)\}$ and $\alpha_T = \beta_T$.

In general, however, expression (6) does not correspond to an MSM in a transparent form. For example, if a logit link is used in model (5), then equation (6) is generally not a logistic regression model. The issue is also known as non-collapsibility.[24]

To avoid imposing implicit constraints on $\mu(t)$, we also consider the following generalized additive model (GAM)

$$m(t, x; \beta) = \psi\{\beta_T(t) + \beta'_X x + \beta'_{TX}(tx)\}, \tag{8}$$

where $\beta_T(t)$ is a free-ranging smooth function. The linear form of $\beta_X'x + \beta_{TX}'(tx)$ is not important here; some elements may be replaced by unspecified smooth functions. What is important is for the model to be identifiable and to depend on t in a flexible way for any given x. It is straightforward to show that, for any range-appropriate smooth function $\mu(t)$ and any given $(\beta_X, \beta_{TX}, F_X)$, where F_X denotes the marginal distribution function of X, condition (4) is satisfied by a unique function $\beta_T(t)$ which is necessarily smooth. Lemma 1 in Appendix gives a precise statement of the result followed by a proof. Thus, model (8) does not impose any stringent constraints on $\mu(t)$.

Model (5) can be fitted using standard techniques such as (iteratively reweighted) least squares, and model (8) using the local scoring algorithm.[25] The local scoring algorithm consists of two nested loops, the inner loop being a weighted backfitting procedure to fit an additive model for an adjusted dependent variable, and the outer loop updating the adjusted dependent variable and the weight. The algorithm can be implemented with any smoother for estimating β_T (e.g. smoothing spline, local regression) as long as it is consistent in the usual setting. Suppose the appropriate algorithm is chosen, and denote the resulting estimate by $\widehat{\beta}$. Then, expression (4) suggests that $\mu(t)$ can be estimated by

$$\widehat{\mu}_{\mathrm{OR}}(t) = \frac{1}{n}\sum_{i=1}^{n} m(t, X_i; \widehat{\beta}),$$

with the subscript OR denoting ordinary regression.

In the case of model (5), standard asymptotic arguments[26] can be used to show that $\sqrt{n}(\widehat{\mu}_{\mathrm{OR}} - \mu)$ converges to a zero-mean Gaussian process with a covariance structure that can be easily derived. In the special cases corresponding to equation (7), we have $\widehat{\mu}_{\mathrm{OR}}(t) = \mu(t; \widehat{\alpha})$ with $\widehat{\alpha}$ obtained by substituting parameter estimates and sample averages in the defining expressions. In Case A, inference on α and hence $\mu(t;\alpha)$ can be based directly on the variance estimate for $\widehat{\beta}$. In Case B, a closed-form variance estimate can be obtained using the following formulas. Let $H = (H_1, H_T, H_X')'$ denote the influence function for $\widehat{\beta}$; these are functions of (X, T, Y) such that $\sqrt{n}(\widehat{\beta} - \beta) = n^{-1/2}\sum_{i=1}^{n} H_i + o_p(1)$, where $H_i = H(X_i, T_i, Y_i)$. For example, if $\widehat{\beta}$ is the maximum likelihood estimator based on a Poisson regression model (with the log link without an interaction between T and X), then $H = I_\beta^{-1}[Y - \exp\{(1, T, X')\beta\}](1, T, X')'$ with

$$I_\beta = \mathrm{E}\big[\exp\{(1, T, X')\beta\}(1, T, X')'(1, T, X')\big],$$

being the Fisher information for β. For any H, the influence function for $\widehat{\alpha}$ is given by $G = (G_1, G_T)'$ with $G_T = H_T$ and

$$G_1 = H_1 + \frac{\exp(\beta_X'X) + \mathrm{E}\{\exp(\beta_X'X)X'\}H_X}{\mathrm{E}\{\exp(\beta_X'X)\}},$$

and the associated variance estimator for $\widehat{\alpha}$ is $n^{-2}\sum \widehat{G}_i\widehat{G}_i'$, where $\widehat{G}_i$ is just $G_i = G(X_i, T_i, Y_i)$ with β replaced by $\widehat{\beta}$ and expectations by sample averages in the defining expressions. In the case of model (8), the non-parametric estimate $\widehat{\mu}_{\mathrm{OR}}(t)$ cannot attain the parametric rate, but a bootstrap procedure can be used to make inference on $\mu(t)$.

3.2 (Augmented) IPW

The IPW approach involves two models: an MSM $\mu(t; \alpha)$, which may be parametric or non-parametric, and a propensity function model $r(t|x; \gamma)$. One possible choice for the MSM is given

by expression (7). The propensity function model is just a parametric model for the conditional distribution of T given X. If T is continuous, it is natural to consider a normal linear model or a Box–Cox-transformed linear model[27] and, if necessary, suitable transformations for some or all elements of X. If T follows a mixed distribution, with both discrete and continuous components, it will be necessary to model each component separately.[28,29]

For a given propensity function model, the parameter γ can be estimated by maximizing the likelihood $\prod_{i=1}^{n} r(T_i|X_i; \gamma)$, and the resulting estimate will be denoted by $\widehat{\gamma}$. Extending the standard inverse probability weighting approach,[4,30] Robins et al.[19] propose to estimate α by fitting the MSM to $\{(T_i, Y_i) : i = 1, \ldots, n\}$ with each subject weighted by $W(T_i)/r(T_i|X_i; \widehat{\gamma})$, where $W(t) > 0$ is chosen to stabilize the weight when $r(T_i|X_i; \widehat{\gamma})$ in the denominator is very small for some subjects. The function W can be data driven but is assumed to converge to a deterministic function w such that $\int w(t)\mathrm{d}\nu(t) = 1$. This IPW approach is justified by the fact that

$$\mathrm{E}\left[\frac{w(T)\{Y - \mu(T; \alpha)\}}{r(T|X; \gamma)}\middle| T\right] \equiv 0, \tag{9}$$

at the true values of α and γ. A proof of this identity, which is not immediate from standard arguments assuming a discrete T, is given in Appendix. Equation (9) holds for an arbitrary function W; however, extremely large weights for some subjects can result in inefficient or erratic estimates. A reasonable choice for W, suggested by Robins et al.,[19] is an estimate of f_T, the marginal density function of T, which can be specified and estimated similarly to the propensity function except that no covariates are involved. We do not propose other choices of W, but keep the notation W to emphasize the fact that f_T need not be specified correctly or estimated well. Robins et al.[19] also suggest basing inference on the sandwich variance estimate. The IPW method has been used in some real-world applications.[31–33]

Further insights into the IPW approach can be gained from the following heuristic argument. Recall that direct (unweighted) fitting of the MSM would be invalid due to confounding. Weighting each subject by the inverse propensity function has the effect of creating a pseudo-population that can be shown to be free of confounding. To avoid confusion, we will use the superscript * to denote random variables in the pseudo-population. Thus, the joint density of (X^*, T^*) is proportional to the joint density of (X, T) multiplied by the (asymptotic) weight and can be written as

$$f_{X^*,T^*}(x, t) = \frac{w(t)}{r(t|x)} f_{X,T}(x, t) = \frac{w(t)}{r(t|x)} f_X(x) r(t|x) = w(t) f_X(x), \tag{10}$$

where f denotes a generic probability density or mass function with the subscript indicating the random variable(s) concerned. Because w is assumed to integrate to 1, the right-hand side is indeed a joint density and, in fact, $w(t)$ is the density of T^*. Equation (10) shows that X^* and T^* are independent and that X^* is distributed as X. If $w = f_T$, then T^* is distributed as T as well. Thus, the main effect of IPW is to remove the dependence of T on X and hence the confounding. Furthermore, because the weighting involves only (X, T), the conditional distribution $(Z^*|X^*, T^*)$ is the same as the unstarred version, where Z^* may be Y^* or $Y^*(t)$ for any t.

For a parametric MSM (i.e. finite dimensional α), the IPW method can be improved under an AIPW approach, which projects an IPW estimating function into the orthogonal complement of the nuisance tangent spact.[10,34] Let $e(Y, T; \alpha)$ denote an estimating function for α that would be used in the absence of confounding; this is typically given by $e(Y, T; \alpha) = \{Y - \mu(T; \alpha)\}h(T)$ for a

vector-valued function h of the same dimension as α. The AIPW approach is based on the following estimating function

$$D(Y, T, X; \alpha) = \frac{w(T)\{e(Y, T; \alpha) - Q(T, X)\}}{r(T|X; \gamma)} + \int w(t)Q(t, X)\mathrm{d}\nu(t),$$

which has mean 0 at the true value of α if either $r(T|X;\gamma) = r(T|X)$ or

$$Q(T, X) = \mathrm{E}\{e(Y, T; \alpha)|T, X\} = \{\mathrm{E}(Y|T, X) - \mu(T; \alpha)\}h(T),$$

with α set at the true value, under suitable regularity conditions.[10,34] Note that the desired $Q(T, X)$ is directly related to the m-model: $m(T, X; \beta) = \mathrm{E}(Y|T, X)$. Denote by $\widehat{D}(Y, T, X; \alpha)$ an estimate of $D(Y, T, X; \alpha)$ with w replaced by W, γ by $\widehat{\gamma}$, and $Q(T, X)$ by a suitable estimate based on the m-model. Then, the AIPW estimate of α is obtained by solving the equation $\sum_{i=1}^{n} \widehat{D}(Y_i, T_i, X_i; \alpha) = 0$. Under regularity conditions, the AIPW estimate is consistent and asymptotically normal under correct specification of either the r-model or the m-model, but not necessarily both. When both models are correctly specified, the AIPW estimator is efficient in the semiparametric sense.[35] In order to take advantage of the double robustness, the m-model needs to be compatible with the μ-model in the sense of (4). In light of the discussion surrounding (7), it appears that the double robustness of the AIPW method is largely limited to the special cases (A and B) considered in Section 3.1. On the other hand, the AIPW method can be more efficient than the IPW method even with an incorrect m-model.

3.3 Stratification

Joffe and Rosenbaum[21] and Imai and Van Dyk[20] assume that the propensity function $r(t|x)$ depends on x only through $u(x)$ for some function u, that is

$$r(t|x_1) = r(t|x_2) \quad \text{whenever} \quad u(x_1) = u(x_2). \tag{11}$$

An important implication of this assumption is that treatment assignment is random within each sub-population defined by $U = u(X)$,[20] that is

$$T \perp Y(t)|U, \quad t \in \mathcal{T}.$$

This further implies that

$$\mathrm{E}(Y|T = t, U) = \mathrm{E}\{Y(t)|T = t, U\} = \mathrm{E}\{Y(t)|U\}, \tag{12}$$

so that the dose-response relationship in each sub-population can be assessed by directly regressing Y on T as in a randomized experiment. The overall does-response relationship, $\mu(t) = \mathrm{E}[\mathrm{E}\{Y(t)|U\}]$, can then be recovered by averaging across sub-populations.

This strategy can be implemented via stratification if the dimension of $u(x)$ is very low, say one or two. This dimensional requirement is met for some choices of the propensity function model $r(t|x; \gamma)$, under which $u(x) = u(x; \gamma)$ may depend on all or part of γ. For example, if we assume that $(T|X = x)$ is distributed as $N(x'\lambda, \sigma^2)$, then $u(x; \lambda) = x'\lambda$ is one dimensional. Further examples for categorical and ordinal treatments are discussed by Imai and Van Dyk.[20] Suppose the dimension of $u(x)$ is

sufficiently low, and let the range of $u(x)$ be stratified as $\mathcal{S}_k$, $k = 1, \ldots, K$. Experience with propensity score stratification for a binary treatment suggests that, if $u(x)$ is one dimensional, it may be reasonable to define the strata using quintiles of U. If each stratum is reasonably homogeneous, it makes sense to define $\mu_k(t) = \mathrm{E}\{Y(t)|U \in \mathcal{S}_k\}$, the dose–response relationship in the kth stratum. Write $\widehat{U}_i = u(X_i; \widehat{\gamma})$ and $\widehat{\mathcal{S}}_k$ for an estimate of $\mathcal{S}_k$ based on the $\widehat{U}_i$. Motivated by equation (12), an estimate of $\mu_k(t)$, say $\widehat{\mu}_k(t)$, can be obtained by regressing Y_i on T_i in $\widehat{\mathcal{S}}_k$, parametrically (by specifying a stratum-specific structural model) or non-parametrically. The overall dose–response relationship $\mu(t)$ can then be estimated by

$$\widehat{\mu}_{\mathrm{S}}(t) = \sum_{k=1}^{K} p_k \widehat{\mu}_k(t),$$

where $p_k = n^{-1} \sum_{i=1}^{n} I\{\widehat{U}_i \in \widehat{\mathcal{S}}_k\}$ is the proportion of subjects falling into the kth stratum. This stratified estimate is generally inconsistent (due to imperfect homogeneity of the strata) but the asymptotic bias should be small with an approximately correct propensity function model and a set of approximately homogeneous strata.

3.4 Summary

Other than the AIPW method described in Section 3.2, each of the aforementioned methods is based on either an m-model or an r-model, and its consistency depends on the relevant model being correctly specified. The AIPW method, which involves both models, is doubly robust and locally efficient. However, the double robustness of the AIPW method appears limited to Cases A and B, as mentioned earlier, and is not available for a binary outcome (with a range-approriate link function). Also, it can be challenging to implement the AIPW method, which involves an integral in the estimating function.

4 Proposed methods

We now propose two methods that are similar to the AIPW method in the sense of double robustness, but easier to implement and more generally applicable (beyond Cases A and B). As mentioned earlier, double robustness requires that the m-model be compatible with the MSM in the sense of (4). This requirement is met in Cases A and B with equation (7) as the MSM and the GLM (5) as the m-model. Beyond those cases, one might be tempted to simply assume, as an MSM, that condition (4) holds for some β, even when the m-model is misspecified. However, with a GLM (5) that does not satisfy equation (7), equation (4) as an assumption is not straightforward to interpret. While it may be acceptable to make assumptions about $\mu(t)$, it is always important for the investigator to understand what is assumed. To address that interpretability problem, we relax the m-model and work with the GAM (8), which does not constrain $\mu(t)$ through equation (4). Thus, beyond Cases A and B, we work with a non-parametric MSM in conjunction with an m-model given by equation (8), which covers a broad range of practical situations such as logistic regression for a binary outcome.

4.1 Weighted regression

Our first proposal is a WR method which is a simple combination of IPW and OR. The idea is to fit an outcome regression model $m(t, x; \beta)$, which may be a GLM given by equation (5) or a GAM

given by equation (8), using the same IPW as in Section 3.2. Denote by $\widehat{\beta}^*$ the resulting estimate of β. Then, the proposed estimate of $\mu(t)$ is given by

$$\widehat{\mu}_{\mathrm{WR}}(t) = \frac{1}{n}\sum_{i=1}^{n} m(t, X_i; \widehat{\beta}^*). \tag{13}$$

This approach has been considered by Hirano and Imbens,[36] Kang and Schaffer[15] and Joffe[37] for a binary exposure but not for a quantitative exposure. Under appropriate conditions (to be specified later), the WR method is doubly robust in the sense that $\mu(t)$ is estimated consistently assuming correct specification of either the m-model or the r-model but not necessarily both. The consistency of $\widehat{\mu}_{\mathrm{WR}}(t)$ under a correctly specified m-model is easy to establish, because the IPW is a function of (T, X) (covariates in the m-model). This is analogous to weighted least squares, which affects the efficiency but not the consistency of the estimate. The effect of IPW on the efficiency of estimating β in the m-model will be studied in simulations. The consistency of $\widehat{\mu}_{\mathrm{WR}}(t)$ based on a correctly specified r-model is not as straightforward to see, as it depends on the structure of the m-model and the method of estimation.

In the special cases corresponding to equation (7), we obtain $\widehat{\beta}^*$ by solving the weighted likelihood equation

$$\sum_{i=1}^{n} r(T_i|X_i; \widehat{\gamma})^{-1} W(T_i)[Y_i - \psi\{(1, T_i, X_i', T_iX_i')\beta\}](1, T_i, X_i', T_iX_i')' = 0. \tag{14}$$

In these cases, expression (13) reduces to $\widehat{\mu}_{\mathrm{WR}}(t) = \mu(t; \widehat{\alpha}^*)$ with $\widehat{\alpha}^*$ obtained by substituting $\widehat{\beta}^*$ and the empirical distribution of X in the expressions defining α. We show in Appendix that $\widehat{\alpha}^*$ is consistent for α under correct specification of the r-model and the MSM (7). Asymptotic normality of $\widehat{\alpha}^*$ follows from standard arguments. In Case A, a variance estimate for $\widehat{\alpha}^*$ can be obtained as a sub-matrix of a variance estimate for $\widehat{\beta}^*$. In Case B, the asymptotic variance of $\widehat{\alpha}^*$ can be derived using the formulas given at the end of Section 3.1, with β replaced by β^*, the probability limit of $\widehat{\beta}^*$.

For the GAM (8), the local scoring algorithm will be used with the same IPW as in Section 3.2. The consistency and convergence properties of the local scoring algorithm have been studied by Buja et al.[38] for the identity link and seem less well understood for GAMs. Here, we assume that the local scoring algorithm is consistent when model (8) is correctly specified and convergent otherwise. Denote by β^* the probability limit of $\widehat{\beta}^*$. Assuming a canonical link such as the logit link for binary data, we show in Appendix that β^* satisfies condition (4) under a correct r-model, even if the m-model is misspecified. Therefore, the WR estimate $\widehat{\mu}_{\mathrm{WR}}(t)$ is doubly robust in the present situation. The associated inference can be performed using a bootstrap procedure.

4.2 Stratified regression

We also propose a stratified regression approach that combines stratification with OR. Under this approach, we specify and estimate a propensity function model $r(t|x; \gamma)$ which satisfies equation (11) for a low-dimensional $u(x)$, and use the estimates $\widehat{U}_i$ to stratify the sample as in Section 3.3. We also specify an outcome regression model $m(t, x; \beta)$ and fit it separately in each stratum. Let $\widetilde{\beta}_k$ denote the resulting estimate of β in the kth stratum. We then estimate $\mu_k(t)$ by

$$\widetilde{\mu}_k(t) = \sum_{i=1}^{n} I\{\widehat{U}_i \in \widehat{\mathcal{S}}_k\} m(t, X_i; \widetilde{\beta}_k) \Big/ \sum_{i=1}^{n} I\{\widehat{U}_i \in \widehat{\mathcal{S}}_k\}.$$

The proposed estimate of $\mu(t)$ is given by

$$\widehat{\mu}_{\mathrm{SR}}(t) = \sum_{k=1}^{K} p_k \widetilde{\mu}_k(t),$$

with p_k defined in Section 3.3.

We recognize that Imai and Van Dyk[20] also consider fitting an m-model in each stratum. However, their proposal is to summarize the estimates $\widetilde{\beta}_k$ through the weighted average $\sum_{k=1}^{K} p_k \widetilde{\beta}_k$. As we shall see, except in Case A, this weighted average is not directly related to our inferential target $\mu(t)$. Moreover, Imai and Van Dyk[20] do not consider the possibility of non-collapsibility, which makes it difficult to assign causal interpretations to their estimates in certain situations such as logistic regression for binary data.

The stratified regression method we propose is approximately doubly robust in the sense that $\widehat{\mu}_{\mathrm{SR}}(t)$ is consistent under the m-model and approximately consistent (like $\widehat{\mu}_{\mathrm{S}}(t)$) under the r-model, under conditions similar to those in Section 4.1. To understand the consistency of $\widehat{\mu}_{\mathrm{SR}}(t)$ under the m-model, we note that the strata are defined by X, which is a covariate vector in the m-model. If the m-model is correct, it is correct in each stratum. With K fixed and n increasing, each $\widetilde{\beta}_k$ converges to the true value of β, and therefore $\widetilde{\mu}_k(t)$ converges to $\mu_k(t) = \mathrm{E}\{m(t, X; \beta) | U \in \mathcal{S}_k\}$. It follows that $\widehat{\mu}_{\mathrm{SR}}(t)$ is consistent for $\mu(t)$. Relative to the OR approach, there may be a loss of efficiency here due to separate estimation of β in each stratum, which will be assessed in simulations. Now, assume that the r-model holds and that each stratum is sufficiently homogeneous. Then, we can treat each stratum as a pseudo-randomized experiment that is largely free of confounding,[20] much like the pseudo-population resulting from IPW (Section 3.2). From this perspective, the question of how to insure consistency under the r-model is similar to the one addressed in Section 4.1. This connection allows us to draw upon the insights developed in Section 4.1, as we now demonstrate.

Consider first the special cases corresponding to equation (7). For each k, we obtain $\widetilde{\beta}_k$ by solving the estimating equation

$$\sum_{i=1}^{n} I(\widehat{U}_i \in \widehat{\mathcal{S}}_k)[Y_i - \psi\{(1, T_i, X_i', T_i X_i')\beta\}](1, T_i, X_i', T_i X_i')' = 0,$$

which resembles equation (14) except that the regression is stratified rather than weighted. In these special cases, we have $\widetilde{\mu}_k(t) = \psi\{(1, t)\widetilde{\alpha}_k\}$ with $\widetilde{\alpha}_k$ obtained by substituting $\widetilde{\beta}_k$ and the empirical distribution of X in the kth stratum. Assuming that $\mu_k(t) = \psi\{(1, t)\alpha_k\}$ for some α_k, we show in Appendix that $\widetilde{\alpha}_k$ is approximately consistent for α_k under correct specification of the r-model. This implies that each $\widetilde{\mu}_k(t)$ is approximately consistent for $\mu_k(t)$ and so is $\widehat{\mu}_{\mathrm{SR}}(t)$ for $\mu(t)$. Now, let us focus on Case A, where we have $\mu(t) = (1, t)\alpha$ with $\alpha = \sum_{k=1}^{K} \pi_k \alpha_k$ and $\pi_k = \mathrm{P}(U \in \mathcal{S}_k)$. It is easy to see that, in this case, $\widehat{\mu}_{\mathrm{SR}}(t) = (1, t)\widetilde{\alpha}$ with $\widetilde{\alpha} = \sum_{k=1}^{K} p_k \widetilde{\alpha}_k$. Considering that $\widetilde{\alpha}_k$ is a sub-vector of $\widetilde{\beta}_k$ if X is centered in each stratum, the stratified regression method is equivalent to the proposal of Imai and Van Dyk[20] in this case. A simple variance estimate for $\widetilde{\alpha}$ is given by $\sum_{k=1}^{K} p_k^2 \widetilde{\Sigma}_k$, where $\widetilde{\Sigma}_k$ is the variance estimate for $\widetilde{\alpha}_k$, which can be obtained as a sub-matrix of the variance estimate for $\widetilde{\beta}_k$. In Case B, we have $\mu(t) = \sum_{k=1}^{K} \pi_k \exp\{(1, t)\alpha_k\}$ and $\widehat{\mu}_{\mathrm{SR}}(t) = \sum_{k=1}^{K} p_k \exp\{(1, t)\widetilde{\alpha}_k\}$, and the stratified regression method is different from the proposal of Imai and Van Dyk,[20] which is now inconsistent for estimating $\mu(t)$. The formulas at the end of Section 3.1 can again be used to derive the asymptotic variances of $\widetilde{\alpha}_k$ and $\widehat{\mu}_{\mathrm{SR}}(t)$.

Beyond these special cases, we propose to fit a GAM like (8) in each stratum using the local scoring algorithm. With a canonical link, the resulting $\widetilde{\mu}_k(t)$ and $\widehat{\mu}_{\mathrm{SR}}(t)$ are approximately consistent

under the *r*-model. This is easy to see by adapting and combining the arguments in Appendix; a formal proof is omitted. A bootstrap procedure can be used to make inference in this case.

Theoretically, this stratified regression method is not as appealing as the AIPW and WR methods. However, the latter methods, like the IPW method, can be numerically unstable when the estimated propensity function is very small for some subjects. The problem is well documented for doubly robust methods involving inverse probability weighting in the literature for binary treatments.[6,15] Like the stratification method, the stratified regression method can be expected to be more stable numerically.

5 Numerical results

5.1 An obstetric example

We now use the methods described in Sections 3 and 4 to assess the effect of a mother's pre-pregnancy BMI on the infant's birth weight. The BMI is a heuristic proxy for human body fat based on an individual's weight and height, and is often used to define obesity, a major public health concern in the United States and worldwide. As a risk factor, the BMI is often dichotomized into an indicator of obesity in epidemiologic research. While convenient, such dichotomization results in a loss of information and possibly an oversimplification of the causal effect of interest. The methods described and proposed in this article allow us to make use of all relevant information in the original BMI and better characterize its effect on the birth weight.

Our research question arose from a large obstetric study known as the Consortium on Safe Labor (CSL).[39] The CSL is a retrospective observational study conducted by the National Institutes of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, in collaboration with 12 institutions across the United States. The goal was to collect comprehensive information on contemporary labor and delivery practice in the United States population. Participating institutions extracted detailed information from their electronic medical records on maternal demographic characteristics, medical history, reproductive and prenatal history, labor and delivery summary, and postpartum and newborn information.

Our analysis is based on 5194 Caucasian women in the CSL with complete information on the outcome (birth weight in grams), the exposure (BMI in kg/cm^2), and all relevant covariates (identified prospectively as maternal age in years, parity, smoking, and diabetes). The outcome regression model we use is a linear model with four linear terms (BMI, maternal age, parity, and smoking) and no interactions. Our propensity function model is also a linear model (assuming normality) which includes maternal age, parity, smoking, and diabetes, as well as interactions of diabetes with maternal age and parity. Both models were constructed using a systematic variable selection procedure, starting with univariate analyses and then considering interactions among the selected covariates. Both models appear to fit the data well.

Table 1 presents the results of estimating the intercept (α_1) and the slope (α_T) in a linear MSM as in Case A. The results are obtained using the methods described in Sections 3 and 4 as well as a naive approach that simply performs a linear regression analysis of birth weight on BMI without adjusting for confounders. For the methods that involve weighting, $W(t)$ is taken to be the marginal density of T estimated under a normal model. For the AIPW method, we adopt the implementation suggested by Neugebauer and Van Der Laan,[34] which in the present context amounts to fitting the *m*-model twice, first without weighting and then with IPW while holding β_X and β_{TX} at the estimates from the first (i.e. unweighted) analysis. Note that assumption (11) is satisfied under the chosen *r*-model, with $u(x)$ being the linear predictor in the *r*-model. The methods involving stratification are implemented with five strata, defined using quintiles of the $\widehat{U}_i$. The results in Table 1 consist of PEs, ASEs based

Table 1. Analysis of the CSL data: point estimates (PEs), analytic standard errors (ASEs), and bootstrap standard errors (BSEs) based on 1000 bootstrap samples, for $\alpha = (\alpha_1, \alpha_T)'$ in a linear MSM: $\mu(t; \alpha) = \alpha_1 + \alpha_T t$.

	PE		ASE		BSE	
Method	α_1	α_T	α_1	α_T	α_1	α_T
NV	3283	9.77	27	1.11	27	1.11
OR	3285	9.69	27	1.10	27	1.09
IPW	3285	9.69	32	1.24	30	1.24
AIPW	3287	9.59	30	1.23	30	1.23
S	3292	9.44	27	1.12	27	1.12
WR	3287	9.59	30	1.23	30	1.23
SR	3288	9.58	27	1.11	27	1.11

CSL: Consortium on safe labor; MSM: Marginal structural model; NV: Naive; OR: Ordinary regression; IPW: Inverse propensity weighting; AIPW: Augmented inverse propensity weighting; S: Stratification; WR: Weighted regression; and SR: Stratified regression.
The results are based on the naive method, OR, IPW, AIPW, stratification, WR, and stratified regression.

on sandwich variance estimates, as well as BSEs based on 1000 bootstrap samples. Considering the sampling variability, the results are fairly consistent between the different methods. Also, remarkable consistency is observed between the ASEs and the BSEs for each method. Overall, Table 1 indicates clearly that a unit increase in the BMI leads to an increase in the birth weight of 9–10 g (assuming, of course, that assumptions (1) and (2) are met, that the linear MSM holds, and that at least one of the m- and r-models is nearly correct).

The results of analyzing a single sample may be arbitrary and inadequate for method comparison; so, we compare the same methods in simulation experiments mimicking the CSL study. Specifically, we focus on the same set of 5194 Caucasian women in the CSL with their original covariate values, and generate values of (T, Y) using the r- and m-models described earlier with parameter values estimated from the previous analyses. The working m- and r-models may or may not be the same as the models for data generation. For both models, misspecification results from omitting an important covariate (smoking). This is motivated by the fact that investigators are primarily concerned about omitting important confounders in analyzing the CSL data. If all such variables are identified and measured, plausible models could be constructed using conventional model-building techniques and serious misspecifications could be detected given the large sample size. In contrast, it is much more difficult, if not impossible, to insure that all important covariates have been considered, due to the finiteness of human knowledge at any point in time.

Possible misspecification of the working models gives rise to four scenarios (both models correct, r-model misspecified, m-model misspecified, and both models misspecified). In each scenario, 1000 datasets are generated and analyzed using the same methods as in Table 1 (with ASEs), and the results are summarized in Table 2 in terms of bias, SD, median standard error, and CP (at level 0.95) in estimating the slope (α_T) in the MSM: $\mu(t; \alpha) = \alpha_1 + \alpha_T t$. Somewhat surprisingly, the bias for the naive method is very small, presumably because confounding effects in different directions tend to cancel each other. Specifically, smoking appears to cause a downward bias, while maternal age and parity appear to have the opposite effect. Although the amount of net confounding is small, Table 2 shows that the bias can be reduced by appropriate adjustments under correctly specified m- and/or r-models. The IPW and stratification methods are clearly biased when the r-model is misspecified,

Table 2. Simulations based on the CSL example: bias, standard deviation (SD), median (analytic) standard error (SE), and coverage probability (CP) at level 0.95, for estimating α_T in a linear MSM: $\mu(t;\alpha) = \alpha_1 + \alpha_T t$, using the same methods as in Table 1.

Method	Bias	SD	SE	CP
		Both models correct		
NV	0.16	0.99	1.01	0.96
OR	0.09	0.99	1.00	0.96
IPW	0.10	1.09	1.05	0.95
AIPW	0.10	1.09	1.03	0.95
S	−0.03	1.00	1.01	0.96
WR	0.10	1.09	1.03	0.95
SR	0.09	1.00	1.01	0.96
		Only m-model correct		
NV	0.09	1.01	1.01	0.95
OR	0.01	1.01	1.00	0.95
IPW	−0.46	1.06	1.04	0.93
AIPW	0.00	1.06	1.02	0.95
S	−0.53	1.01	1.01	0.93
WR	0.00	1.06	1.03	0.95
SR	0.01	1.01	1.01	0.95
		Only r-model correct		
NV	0.08	1.03	1.01	0.94
OR	−0.47	1.02	1.01	0.92
IPW	0.01	1.09	1.04	0.95
AIPW	0.01	1.09	1.04	0.95
S	−0.12	1.03	1.01	0.95
WR	0.01	1.09	1.04	0.95
SR	0.01	1.03	1.01	0.95
		Neither model correct		
NV	0.04	1.01	1.01	0.96
OR	−0.51	1.00	1.01	0.92
IPW	−0.49	1.04	1.04	0.93
AIPW	−0.49	1.04	1.03	0.93
S	−0.58	1.00	1.01	0.92
WR	−0.49	1.04	1.03	0.93
SR	−0.51	1.00	1.01	0.93

CSL: Consortium on safe labor; MSM: Marginal structural model; NV: Naive; OR: Ordinary regression; IPW: Inverse propensity weighting; AIPW: Augmented inverse propensity weighting; S: Stratification; WR: Weighted regression; and SR: Stratified regression.
Each entry is based on 1000 replicates.

and so is the OR method when the *m*-model is misspecified. These biases are larger than the bias of the naive method because the misspecification (deleting smoking from the model) results in a larger amount of net confounding. As expected, the AIPW, WR, and stratified regression methods are nearly unbiased when either model is correct. When both models are misspecified, all methods (except the naive one) are severely biased. In all scenarios, the methods that involve weighting

(IPW, AIPW, and WR) tend to be less efficient than the other methods, which are largely equal in efficiency. The ASEs based on sandwich variance estimates appear to work well in this situation, and the CP is usually close to the nominal level unless there is a large amount of bias.

5.2 Additional simulations

In this section, we report two additional sets of simulations, one for a continuous outcome following a linear MSM and one for a binary outcome with a non-parametric MSM. In both sets of simulations, X is a scalar distributed as $N(0, 1)$, and $T = X + \epsilon$ with $\epsilon \sim N(0, 1)$ (independently of X). In the first set of simulations, the outcome is generated as $Y = T + X + \xi$ with $\xi \sim N(0, 1)$ (independently of T and X). Under assumption (1), this implies that $\mu(t) = \alpha_1 + \alpha_T t$ with $\alpha_1 = 0$ and $\alpha_T = 1$. The objective is to estimate α_T. The working r-model is a normal linear regression model with X (correct) or X^3 (incorrect) as the only covariate. The working m-model is a linear regression model with covariate vector (T,X) (correct) or (T, X^3) (incorrect) without an interaction term. Each dataset consists of 1000 subjects, and in each scenario, 1000 datasets are generated and analyzed using the same methods as in Table 2. The results, presented in Table 3, contain some new information relative to Table 2. The naive method is now severely biased, the IPW method has a slight bias even under a correct r-model, and the stratification method appears to have a constant bias regardless of the working models. To understand the last observation, we note that stratification on X^3 is equivalent to stratification on X with appropriately transformed cut-off values. Thus, regardless of the working r-model, we have the same strata based on quintiles of X. One could argue that the design of this simulation study (with X misspecified as X^3) gives an advantage to the methods involving stratification. On the other hand, it does illustrate a robustness property of these methods (i.e. the invariance of strata under monotone transformations). Another phenomenon not observed in Table 2 is that the (analytic and bootstrap) standard errors do not work well for the methods that involve weighting. This may have to do with the fact that the weights here are more variable than in the CSL example. After dividing by the mean weight, the largest weight is 5.3 in the CSL and 32.8 in a random sample of size 1000 in the present situation. The sandwich variance estimate, which treats the weights as fixed, may not be appropriate in the present situation. The BSE is not satisfactory, either, although it seems to work better than the ASE when the r-model is misspecified. Therefore, to obtain confidence intervals for the IPW, AIPW, and WR methods, we also explore a simple bootstrap percentile (BP) method which takes the 2.5th and 97.5th percentiles of the bootstrap distribution. In Table 3, good coverage of a normality-based confidence interval with an ASE is achieved for OR under a correct m-model and for stratified regression when either model is correctly specified. For the AIPW and WR methods, the BP confidence interval appears satisfactory under correct specification of either the m-model or the r-model, while those based on standard errors do not always work well, suggesting that the sampling distribution of the point estimator may be non-normal. The lack of good coverage for the other methods (naive, IPW, and stratification) may be explained by bias from various sources (inconsistency, finite-sample bias, and imperfect stratification).

Table 4 summarizes the second set of simulations, where X and T are generated the same way as before and a binary outcome is generated according to the logistic model: $\mathrm{P}(Y = 1|T, X) = \mathrm{expit}(T - X)$. This corresponds to $\mu(t) = \int \mathrm{expit}(t - x)\mathrm{d}\Phi(x)$, where Φ is the standard normal distribution function. For method comparison, we are interested in estimating $\mu(t_{0.25})$, where t_a is the ath quantile of T, and the interquartile log-odds ratio $\rho = \mathrm{logit}\{\mu(t_{0.75})\} - \mathrm{logit}\{\mu(t_{0.25})\}$. (Because of the data generation mechanism, all methods are unbiased for estimating $\mu(t_{0.5})$, and the estimation of $\mu(t_{0.75})$ is symmetric to the estimation of $\mu(t_{0.25})$.) No parametric assumptions are made about $\mu(t)$, so the AIPW method is not applicable here. The other methods are applicable

Table 3. Additional simulations for a continuous outcome: bias, SD, median standard error (ASE for analytic, BSE for bootstrap with 1000 samples) and CP at level 0.95 based on ASE, BSE or BP, for estimating α_T in a linear MSM: $\mu(t; \alpha) = \alpha_1 + \alpha_T t$, using the same methods as in Table 1.

					CP		
Method	Bias	SD	ASE	BSE	ASE	BSE	BP
			Both models correct				
NV	0.50	0.03	0.03		0.00		
OR	0.00	0.03	0.03		0.95		
IPW	0.06	0.13	0.07	0.06	0.63	0.60	0.60
AIPW	0.00	0.07	0.05	0.05	0.88	0.90	0.93
S	0.09	0.03	0.03		0.22		
WR	0.00	0.07	0.05	0.05	0.89	0.91	0.93
SR	0.00	0.03	0.03		0.94		
			Only m-model correct				
NV	0.50	0.03	0.03		0.00		
OR	0.00	0.03	0.03		0.95		
IPW	0.84	0.41	0.07	0.22	0.02	0.06	0.03
AIPW	0.01	0.18	0.03	0.09	0.46	0.95	0.95
S	0.09	0.03	0.03		0.23		
WR	0.01	0.18	0.04	0.05	0.80	0.86	0.95
SR	0.00	0.03	0.03		0.94		
			Only r-model correct				
NV	0.50	0.03	0.03		0.00		
OR	0.28	0.04	0.03		0.00		
IPW	0.07	0.13	0.07	0.06	0.62	0.60	0.60
AIPW	0.01	0.09	0.06	0.06	0.75	0.90	0.92
S	0.09	0.03	0.03		0.20		
WR	0.01	0.09	0.06	0.05	0.84	0.88	0.92
SR	0.01	0.03	0.03		0.93		
			Neither model correct				
NV	0.50	0.03	0.03		0.00		
OR	0.28	0.04	0.03		0.00		
IPW	0.85	0.42	0.07	0.22	0.01	0.06	0.03
AIPW	0.24	0.15	0.06	0.31	0.29	0.85	0.92
S	0.09	0.03	0.03		0.21		
WR	0.24	0.15	0.03	0.05	0.07	0.12	0.12
SR	0.01	0.03	0.03		0.92		

SD: Standard Deviation; ASE: Analytic standard error; BSE: Bootstrap standard error; CP: Coverage probability; BP: Bootstrap percentile; MSM: Marginal structural model; NV: Naive; OR: Ordinary regression; IPW: Inverse propensity weighting; AIPW: Augmented inverse propensity weighting; S: Stratification; WR: Weighted regression; and SR: Stratified regression. Each entry is based on 1000 replicates.

Table 4. Additional simulations for a binary outcome: bias and SD for estimating $\mu(t_{0.25})$ (response probability at the first quartile of T) and the inter-quartile log-odds ratio ρ (defined in Section 5.2) under a non-parametric MSM, using the same methods as in Table 1 (except AIPW).

	Bias		SD	
Method	$\mu(t_{0.25})$	ρ	$\mu(t_{0.25})$	ρ
	Both models correct			
NV	0.08	−0.74	0.03	0.15
OR	0.00	−0.01	0.02	0.15
IPW	0.01	−0.05	0.04	0.26
S	0.00	−0.03	0.03	0.21
WR	0.00	0.20	0.04	3.50
SR	0.00	0.00	0.03	0.19
	Only m-model correct			
NV	0.08	−0.72	0.02	0.14
OR	0.00	0.01	0.02	0.14
IPW	0.13	−1.37	0.16	3.11
S	0.00	0.00	0.03	0.19
WR	0.00	0.09	0.05	2.07
SR	0.00	0.02	0.03	0.18
	Only r-model correct			
NV	0.08	−0.73	0.02	0.15
OR	0.06	−0.49	0.02	0.15
IPW	0.01	−0.03	0.04	0.26
S	0.00	−0.02	0.03	0.19
WR	0.00	0.00	0.06	0.55
SR	0.00	−0.01	0.03	0.19
	Neither model correct			
NV	0.08	−0.72	0.02	0.15
OR	0.05	−0.48	0.02	0.15
IPW	0.13	−1.34	0.15	2.40
S	0.00	−0.01	0.03	0.19
WR	0.06	−0.34	0.17	9.94
SR	0.00	0.00	0.03	0.18

SD: Standard Deviation; MSM: Marginal structural model; AIPW: Augmented inverse propensity weighting; NV: Naive; OR: Ordinary regression; IPW: Inverse propensity weighting; S: Stratification; WR: Weighted regression; and SR: Stratified regression.
Each entry is based on 1000 replicates.

upon specifying a non-parametric additive component for T when regressing Y on T (and possibly X), implemented using the gam package in R. The working r-model (correct or incorrect) is the same as in the last paragraph. The working m-model is an additive logistic regression model with covariate vector (T, X) (correct) or (T, X^3) (incorrect) without an interaction term. Each dataset consists of 1000 subjects, and in each scenario, 1000 datasets are generated and analyzed. Table 4 compares the different methods in terms of bias and standard deviation, as ASEs are not available in this situation. Although the situation here is very different from Table 3, the results in Table 4 follow a similar pattern. Note that the quantities $\mu(t_{0.25})$ and ρ are on different scales, which should be kept in mind

when interpreting the results. Also, an observed bias should always be judged with the sampling variability in mind. For example, the WR method is not significantly biased when both models are correct; the observed bias merely reflects a large amount of variability.

6 Discussion

We have reviewed the existing methods and proposed two new ones (WR and stratified regression) for causal inference with a quantitative exposure. The WR method is similar to AIPW in the sense of double robustness, but easier to implement (in Case B) and more available (for binary outcomes). The stratified regression method is only approximately doubly robust, but it does have some extra robustness in the sense that valid stratification is possible under an incorrect r-model (Section 5.2). In our simulation experiments (in Case A), the WR method behaves in a manner similar to the AIPW method, while the stratified regression method appears more efficient, especially when the IPW is highly variable. In the latter case, the stratified regression method seems more appealing than the other methods. The stratified regression method does require condition (11) to hold for a low-dimensional function $u(x)$, which is the case in our simulation studies but not in general. For example, if the normal linear regression of T on X is heteroskedastic, it will be necessary to model the heteroskedasticity and include additional terms in $u(x)$ to account for the heteroskedasticity. The situation can be even more difficult if $(T|X)$ follows a non-normal or mixed distribution. In such cases, the WR method may be preferable if the IPW is not too variable.

A practical approach, suggested by a referee, is to perform a sensitivity analysis using all available methods and comparing the results, as we did for the CSL. The inference will be more credible if the results from different methods are similar than if they are dissimilar. In the latter case, one might want to reexamine the working models for their goodness of fit. Even with double robustness, one cannot be too careful in specifying the working m- and r- models. Among other possibilities, loss-based cross-validation[40,41] appears to be a promising approach to model selection in this situation.

It should be noted that the sandwich and bootstrap variance estimates do not work well for the methods that involve weighting when the weight is highly variable (Table 3). We recommend the use of BP confidence intervals in this situation.

With focus on prospectively specified models, we have not discussed data-adaptive models such as the loss-based cross-validation approach[40,41] developed recently by Van Der Laan and colleagues. The latter approach attempts to choose parametric models based on the data, in such a way that, asymptotically, the model selection procedure works as well as if the true distribution is known. This approach may have profound implications on statistical modeling in general. Targeted maximum likelihood learning[14] is another promising approach to causal inference with a quantitative exposure.

Acknowledgments

The authors thank the two anonymous referees for insightful comments that have greatly improved the manuscript. The views expressed in this article are not necessarily those of the US Food and Drug Administration.

Funding

Zhiwei Zhang and Jun Zhang were supported in part by the Intramural Research Program of the National Institutes of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

References

1. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974; **66**: 688–701.
2. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
3. Rosenbaum PR and Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *JASA* 1984; **79**: 516–524.
4. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *JASA* 1994; **89**: 846–866.
5. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**: 2265–2281.
6. Lunceford JK and Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004; **23**: 2937–2960.
7. Lipsitz SR, Ibrahim JG and Zhao LP. Weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *JASA* 1999; **94**: 1147–1160.
8. Scharfstein DO, Rotnitzky A and Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse mopdels (with discussion). *JASA* 1999; **94**: 1096–1146.
9. Robins JM Rotnitzky A Bickel PJ Kwon J Comment on "Inference for semiparametric models: some questions and an answer". by PJ Bickel and J Kwon. *Stat Sin* 2001; **11**: 920–936.
10. Van Der Laan MJ and Robins JM. *Unified methods for censored longitudinal data and causality*. New York: Spring-Verlag, 2003.
11. Bang H and Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**: 962–972.
12. Carpenter JR, Kenward MG and Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *JRSS-A* 2006; **169**: 571–584.
13. Tan Z. A distributional approach for causal inference using propensity scores. *JASA* 2006; **101**: 1619–1637.
14. Van Der Laan MJ and Rubin DB. Targeted maximum likelihood learning. *Int J Biostat* 2006; **2**(1): Article 11.
15. Kang JDY and Schaffer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Stat Sci* 2007; **22**: 523–539.
16. Rubin DB and Van Der Laan MJ. Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Int J Biostat* 2008; **4**: Article 5.
17. Cao W, Tsiatis AA and Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* 2009; **96**: 723–734.
18. Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* 2010; **97**: 661–682.
19. Robins JM, Hernan MA and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–560.
20. Imai K and Van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *JASA* 2004; **99**: 854–866.
21. Joffe MM and Rosenbaum PR. Invited commentary: propensity scores. *Am J Epidemiol* 1999; **150**: 327–333.
22. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**: 706–710.
23. Moore KL, Neugebauer R, Van Der Laan MJ, et al. Causal inference in epidemiological studies with strong confounding. *Statistics in Medicine* 2012; **31**: 1380–1404. (DOI: 10.1002/sim.4469).
24. Greenland S, Robins JM and Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999; **14**: 29–46.
25. Hastie TJ and Tibshirani RJ. *Generalized additive models*. New York: Chapman & Hall/CRC, 1990.
26. Van Der Vaart AW. *Asymptotic statistics*. Cambridge: Cambridge University Press, 1998.
27. Box GEP and Cox DR. An analysis of transformations. *JRSS-B* 1964; **26**: 211–252.
28. Zhou X-H and Tu W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics* 1999; **55**: 645–651.
29. Tian L and Huang J. A two-part model for censored medical cost data. *Stat Med* 2007; **26**: 4273–4292.
30. Horvitz DG and Thompson DJ. A generalization of sampling without replacement from a finite universe. *JASA* 1952; **47**: 663–685.
31. Tager IB, Haight T, Sternfeld B, et al. Effects of physical activity and body composition on functional limitation in the elderly. *Epidemiology* 2004; **15**: 479–493.
32. Haight T, Tager I, Sternfeld B, et al. Effects of body composition and leisure-time physical activity on transitions in physical functioning in the elderly. *Am J Epidemiol* 2005; **162**: 607–617.
33. Van Der Laan M, Haight TJ and Hager IB. Rejoinder to "Hypothetical interventions to define causal effects". *Am J Epidemiol* 2005; **162**: 621–622.
34. Neugebauer R and Van Der Laan J. Why prefer double robust estimators? *J Stat Plan Inference* 2005; **129**: 405–428.
35. Bickel PJ, Klaassen CAJ, Ritov Y, et al. *Efficient and adaptive estimation for semiparametric models*. Baltimore, MD: Johns Hopkins University Press, 1993.
36. Hirano K and Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcomes Res Methodol* 2001; **2**: 259–278.
37. Joffe. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Stat Sci* 2007; **22**: 544–559.
38. Buja A, Hastie T and Tibshirani R. Linear smoothers and additive models (with discussion). *Ann Stat* 1989; **17**: 453–555.
39. Zhang J, Troendle J, Reddy UM, et al. for the Consortium on Safe Labor. Contemporary cesarean delivery practice in the United States. *Am J Obstet Gynecol* 2010; **203**: 326.e1–326.e10.
40. Dudoit S, Van Der Laan MJ, Keles S, et al. *Loss-based estimation with cross-validation: applications to microarray data analysis and motif finding*. Berkeley: Technical Report 137, Division of Biostatistics, University of California, 2003.
41. Wang Y, Bembom, Van Der Laan MJ, et al. Data-adaptive estimation of the treatment-specific mean. *Journal of Statistical Planning and Inference* 2007; **137**: 1871–1887.

Appendix

Proofs

We write f for a generic (conditional) probability density or mass function, and F for a cumulative distribution function, with subscripts indicating the random variable(s) concerned.

Flexibility of model (8) with respect to $\mu(t)$

Lemma 1: *Suppose that ψ is continuous and strictly increasing and that X is bounded. For any function μ that maps into the range of ψ and any given $(\beta_X, \beta_{TX}, F_X)$, equation (4) is satisfied by a unique function β_T. Furthermore, β_T can be taken to be continuously differentiable if μ and ψ are continuously differentiable.*

Proof: Let $(\beta_X, \beta_{TX}, F_X)$ be given. For each t, we define

$$g_t(b) = \mathrm{E}\psi\{b + \beta_X' X + \beta_{TX}'(tX)\} = \int \psi\{b + \beta_X' x + \beta_{TX}'(tx)\}\mathrm{d}F_X(x).$$

It is easy to see that g_t is strictly increasing and continuous; the latter follows from the dominated convergence theorem with a bounded X. Therefore, for each t, there exists a unique b_t such that $g_t(b_t) = \mu(t)$. We set $\beta_T(t) = b_t$; then β_T satisfies equation (4). Now suppose μ and ψ are continuously differentiable. It follows from the dominated convergence theorem that the function

$$h(b, t) = \int \psi\{b + \beta_X' x + \beta_{TX}'(tx)\}\mathrm{d}F_X(x).$$

is continuously differentiable. Because β_T defined above satisfies $h(\beta_T(t), t) = \mu(t)$ for every t, it must also be continuously differentiable by the implicit mapping theorem. □

Proof of (9)

Note that

$$\mathrm{E}\left[\frac{w(T)\{Y - \mu(T;\alpha)\}}{r(T|X;\gamma)}\middle| T = t\right] = \frac{w(t)}{f_T(t)}\left[\mathrm{E}\left\{\frac{f_T(t)Y}{r(t|X;\gamma)}\middle| T = t\right\} - \mu(t;\alpha)\mathrm{E}\left\{\frac{f_T(t)}{r(t|X;\gamma)}\middle| T = t\right\}\right]. \quad (15)$$

At the true value of γ in a correctly specified r-model, we have

$$\begin{aligned}\mathrm{E}\left\{\frac{f_T(t)Y}{r(t|X;\gamma)}\middle| T = t\right\} &= \mathrm{E}\left(\mathrm{E}\left\{\frac{f_T(t)Y}{r(t|X)}\middle| X, T = t\right\}\middle| T = t\right) = \int \mathrm{E}\left\{\frac{f_T(t)Y}{r(t|X)}\middle| X = x, T = t\right\}\mathrm{d}F_{X|T}(x|t)\\ &= \int \frac{f_T(t)\mathrm{E}\{Y(t)|X = x\}}{r(t|x)}\mathrm{d}F_{X|T}(x|t) = \int \mathrm{E}\{Y(t)|X = x\}\mathrm{d}F_X(x)\\ &= \mathrm{E}[\mathrm{E}\{Y(t)|X\}] = \mu(t), \end{aligned} \quad (16)$$

where the third step follows from assumption (1) and the next from assumption (2) together with the fact that

$$\mathrm{d}F_{X|T}(x|t)/\mathrm{d}F_X(x) = r(t|x)/f_T(t).$$

Similarly, it can be shown that

$$\mathrm{E}\left\{\frac{f_T(t)}{r(t|X;\gamma)}\middle| T = t\right\} = 1. \quad (17)$$

The proof is complete upon substituting (16) and (17) into (15) and noting that $\mu(t) = \mu(t;\alpha)$ at the true value of α.

Consistency of $\widehat{\alpha}^*$ in Section 4.1 under the r*-model*

Denote by β^* the probability limit of $\widehat{\beta}^*$, which solves the equation

$$\mathrm{E}\big(r(T|X;\gamma)^{-1}w(T)[Y-\psi\{(1,T,X',TX')\beta^*\}](1,T,X',TX')'\big)=0.$$

Focus on the first two elements of the above vector equation, and write

$$\begin{aligned}0&=\mathrm{E}\big(r(T|X;\gamma)^{-1}w(T)[Y-\psi\{(1,T,X',TX')\beta^*\}](1,T)'\big)\\&=\mathrm{E}\big\{\mathrm{E}\big(r(T|X;\gamma)^{-1}w(T)[Y-\psi\{(1,T,X',TX')\beta^*\}](1,T)'|T\big)\big\}\\&=\mathrm{E}\big\{\big(f_T(T)^{-1}w(T)\mu(T)-\mathrm{E}\big[r(T|X;\gamma)^{-1}w(T)\psi\{(1,T,X',TX')\beta^*\}|T\big]\big)(1,T)'\big\},\end{aligned}\tag{18}$$

where the last step follows from (16). It can be shown as in (16) that

$$\mathrm{E}\big[r(T|X;\gamma)^{-1}w(T)\psi\{(1,T,X',TX')\beta^*\}|T=t\big]=f_T(t)^{-1}w(t)\int\psi\{(1,t,x',tx')\beta^*\}\mathrm{d}F_X(x).$$

Substituting this into (18) shows that

$$\begin{aligned}0&=\mathrm{E}\left(f_T(T)^{-1}w(T)\left[\mu(T)-\int\psi\{(1,T,x',Tx')\beta^*\}\mathrm{d}F_X(x)\right](1,T)'\right)\\&=\mathrm{E}\big[f_T(T)^{-1}w(T)\{\psi(\alpha_1+\alpha_T T)-\psi(\alpha_1^*+\alpha_T^*T)\}(1,T)'\big],\end{aligned}\tag{19}$$

where $\alpha^*=(\alpha_1^*,\alpha_T^*)'$ is the probability limit of $\widehat{\alpha}^*$. By appealing to the structure of ψ (identity or log), it is straightforward to deduce from (19) that α^* equals the true value of α. More generally, this follows from the identifiability of α in a well-behaved μ-model in the absence of confounding.

Consistency of $\widehat{m}_{\mathrm{WR}}(t)$ based on model (8) under the r*-model*

In the local scoring algorithm, given the current estimate $\beta^{(j)}=\big(\beta_T^{(j)},\beta_X^{(j)'},\beta_{TX}^{(j)'}\big)'$, the updated estimate $\beta_T^{(j+1)}$ is obtained by smoothing over T_i the adjusted dependent variable

$$\beta_T^{(j)}(T_i)+\big[Y_i-\psi\big\{\beta_T^{(j)}(T_i)+\beta_X^{(j)'}X_i+\beta_{TX}^{(j)'}(T_iX_i)\big\}\big]/\dot{\psi}\big\{\beta_T^{(j)}(T_i)+\beta_X^{(j)'}X_i+\beta_{TX}^{(j)'}(T_iX_i)\big\},$$

where $\dot{\psi}$ denotes the derivative of ψ, with weight given by

$$W(T_i)\dot{\psi}\big\{\beta_T^{(j)}(T_i)+\beta_X^{(j)'}X_i+\beta_{TX}^{(j)'}(T_iX_i)\big\}^2/\big\{r(T_i|X_i;\widehat{\gamma})V_i^{(j)}\big\},$$

where $V_i^{(j)}$ is the model-based variance of $(Y_i|T_i,\ X_i)$ evaluated at $\beta^{(j)}$. Setting $\beta^{(j+1)}=\beta^{(j)}$ (at convergence) and letting $n\to\infty$, we see that the limit $\beta^*=(\beta_T^*,\beta_X^{*\prime},\beta_{TX}^{*\prime})'$ satisfies

$$\begin{aligned}0&\equiv\mathrm{E}\left(\frac{w(T)\dot{\psi}\{\beta_T^*(T)+\beta_X^{*\prime}X+\beta_{TX}^{*\prime}(TX)\}^2}{r(T|X)V^*}\times\left[\beta_T(T)+\frac{Y_i-\psi\{\beta_T^*(T)+\beta_X^{*\prime}X+\beta_{TX}^{*\prime}(TX)\}}{\dot{\psi}\{\beta_T^*(T)+\beta_X^{*\prime}X+\beta_{TX}^{*\prime}(TX)\}}-\beta_T(T)\right]\middle|T\right)\\&=\mathrm{E}\left(\frac{w(T)\dot{\psi}\{\beta_T^*(T)+\beta_X^{*\prime}X+\beta_{TX}^{*\prime}(TX)\}[Y_i-\psi\{\beta_T^*(T)+\beta_X^{*\prime}X+\beta_{TX}^{*\prime}(TX)\}]}{r(T|X)V^*}\middle|T\right),\end{aligned}$$

where V^* is the model-based variance of $(Y|T, X)$ evaluated at β^*. For a canonical link, $\dot{\psi}\{\beta_T^*(T) + \beta_X^{*\prime} X + \beta_{TX}^{*\prime}(TX)\}/V^*$ is a constant, and we have, for each t

$$\begin{aligned} 0 &= \mathrm{E}\big(r(T|X)^{-1} w(T)[Y_i - \psi\{\beta_T^*(T) + \beta_X^{*\prime} X + \beta_{TX}^{*\prime}(TX)\}]|T = t\big) \\ &= f_T(t)^{-1} w(t)\big(\mu(t) - \mathrm{E}\big[\psi\{\beta_T^*(t) + \beta_X^{*\prime} X + \beta_{TX}^{*\prime}(tX)\}\big]\big), \end{aligned}$$

where the second step follows from the same arguments used in the previous proofs. Because $w(t) > 0$, this shows that β^* satisfies equation (4).

Approximate consistency of $\widetilde{\alpha}_k$ in Section 4.2 under the r-model

Denote by β_k^* the probability limit of $\widetilde{\beta}_k$, which solves the equation

$$\mathrm{E}\big(I(U \in \mathcal{S}_k)[Y - \psi\{(1, T, X', TX')\beta_k^*\}](1, T, X', TX')'\big) = 0.$$

The first two elements of this vector equation can be rewritten as

$$\begin{aligned} 0 \quad &= \mathrm{E}\big(I(U \in \mathcal{S}_k)[Y - \psi\{(1, T, X', TX')\beta_k^*\}](1, T)'\big) \\ &= \mathrm{P}(U \in \mathcal{S}_k)\mathrm{E}\big([Y - \psi\{(1, T, X', TX')\beta_k^*\}](1, T)'|U \in \mathcal{S}_k\big) \\ &= \mathrm{P}(U \in \mathcal{S}_k)\mathrm{E}\big\{\mathrm{E}\big([Y - \psi\{(1, T, X', TX')\beta_k^*\}](1, T)'|T, U \in \mathcal{S}_k\big)|U \in \mathcal{S}_k\big\}. \end{aligned} \tag{20}$$

Given U, T is conditionally independent of X and the $Y(t)$.[20] Given $U \in \mathcal{S}_k$, the same conditional independence is approximately true if the stratum $\mathcal{S}_k$ is approximately homogeneous. From this it follows that

$$\begin{aligned} &\mathrm{E}\big([Y - \psi\{(1, T, X', TX')\beta_k^*\}](1, t)'|T = t, U \in \mathcal{S}_k\big) \\ &\quad = \mathrm{E}\big([Y(t) - \psi\{(1, t, X', tX')\beta_k^*\}](1, t)'|T = t, U \in \mathcal{S}_k\big) \\ &\quad \approx \left[\mu_k(t) - \int \psi\{(1, t, x', tx')\beta_k^*\}\mathrm{d}F_{X|k}(x)\right](1, t)', \end{aligned}$$

where $F_{X|k}$ denotes the conditional distribution of X given $U \in \mathcal{S}_k$. Substituting this into (20) yields

$$\begin{aligned} 0 &\approx \mathrm{E}\left(\left[\mu_k(T) - \int \psi\{(1, T, x', Tx')\beta_k^*\}\mathrm{d}F_{X|k}(x)\right](1, T)'\middle| U \in \mathcal{S}_k\right) \\ &= \mathrm{E}\big([\psi\{(1, T)\alpha_k\} - \psi\{(1, T)\alpha_k^*\}](1, T)'|U \in \mathcal{S}_k\big), \end{aligned} \tag{21}$$

where α_k^* is the probability limit of $\widetilde{\alpha}_k$. By appealing to the structure of ψ (identity or log), it is straightforward to deduce from (21) that α_k^* equals the true value of α_k.