Estimating Mann–Whitney-Type Causal Effects

Zhiwei Zhang[1,*], Shujie Ma[1], Changyu Shen[2] and Chunling Liu[3]

[1]Department of Statistics, University of California, Riverside, California 92521, USA

[2]Richard A. and Susan F. Smith Center for Outcomes Research in Cardiology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts 02215, USA

[3]Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, PR China

[*]zhiwei.zhang@ucr.edu

Summary

Mann–Whitney-type causal effects are generally applicable to outcome variables with a natural ordering, have been recommended for clinical trials because of their clinical relevance and interpretability, and are particularly useful in analyzing an ordinal composite outcome that combines an original primary outcome with death and possibly treatment discontinuation. In this article, we consider robust and efficient estimation of such causal effects in observational studies and clinical trials. For observational studies, we propose and compare several estimators: regression estimators based on an outcome regression (OR) model or a generalized probabilistic index (GPI) model, an inverse probability weighted estimator based on a propensity score (PS) model, and two doubly robust (DR), locally efficient estimators. One of the DR estimators involves a PS model and an OR model, is consistent and asymptotically normal under the union of the two models, and attains the semiparametric information bound when both models are correct. The other DR estimator has the same properties with the OR model replaced by a GPI model. For clinical trials, we extend an existing augmented estimator based on a GPI model and propose a new one based on an OR

model. The methods are evaluated and compared in simulation experiments, and applied to a clinical trial in cardiology and an observational study in obstetrics.

Key words: causal inference; double robustness; efficient influence function; local efficiency; ordinal outcome; Wilcoxon–Mann–Whitney test

1 Introduction

The causal effect of a treatment or exposure is usually defined using potential outcomes[1]. Consider a binary treatment ($t = 1$ if treated; 0 if untreated), and let $Y(t)$ denote the potential outcome for treatment level t. The causal effect of $t = 1$ versus $t = 0$ is usually assessed by comparing summary measures (such as means and selected quantiles) of the distributions of $Y(1)$ and $Y(0)$[2,3,4,5,6]. Here we consider a different type of effect measures, $\theta = \mathrm{E}\{h(Y_i(1), Y_j(0))\}$, where $h(\cdot, \cdot)$ is a specified function and the subscripts i and j denote two independent subjects. A popular choice for h, which assumes that the outcome values are ordered, is given by

$$h(y_1, y_0) = I(y_1 > y_0) - I(y_1 < y_0), \tag{1}$$

where $I(\cdot)$ is the indicator function; see, for example, Agresti[7], page 58. Related definitions include

$$h(y_1, y_0) = I(y_1 > y_0) + 0.5I(y_1 = y_0), \tag{2}$$

which forms the basis for the rank sum test[8] and the Mann-Whitney statistic[9]. For a continuous outcome, expression (2) is equivalent to $h(y_1, y_0) = I(y_1 > y_0)$, and the corresponding θ is the probability that treatment 1 produces a higher outcome than treatment 0 if the two treatments are applied to two different subjects randomly. Other choices of h are possible. With an appropriate choice of h, θ provides a natural effect measure for ordered categorical

outcomes without requiring a parametric marginal structural model (such as the proportional odds model). Such effect measures have also been recommended for clinical trials with arbitrary (discrete, continuous or mixed) outcomes because of their clinical relevance and interpretability[10,11,12]. They are particularly useful in analyzing an ordinal composite outcome that combines an original primary outcome with death and possibly treatment discontinuation due to adverse events or lack of efficacy[13,14,15]. For such a composite outcome, the mean is undefined unless one can assign numerical values in a meaningful way, and a comparison of quantiles depends on the choice of quantiles and thus cannot be interpreted as an overall effect measure.

In this paper, we consider estimation of θ for a general (specified) function h. Because θ is not a difference of marginal characteristics of $Y(1)$ and $Y(0)$, most of the existing methods for causal inference are not directly applicable to the estimation of θ. In a randomized clinical trial, θ can be estimated using a standard U-statistic, which does not make use of baseline covariate information. In this context, Vermeulen et al.[16] discuss ways to improve the precision in estimating θ (based on (2)) by incorporating covariate information using semiparametric theory[17,18]. They characterize the influence functions of all regular, asymptotically linear estimators of θ, and identify the one with the smallest variance. They further propose an augmented estimator which involves an augmentation term based on a probabilistic index (PI) model[19], attains the minimum variance when the PI model is correct, and remains consistent and asymptotically normal if the PI model is incorrect. For an observational study, Chen et al.[20] consider how to estimate θ (also based on (2)) under the assumption of strongly ignorable treatment assignment[21]. To adjust for confounding, they use kernel smoothing to regress the outcome on the confounders in each treatment group when the number of confounders is small. With a large number of confounders, they apply the same kernel smoothing technique to a parametric estimate of the propensity score (PS;

the conditional probability of being treated given confounders).

The methodological developments in the present paper are mainly for observational studies, although we also discuss clinical trials briefly. For an observational study, we derive the efficient influence function for estimating θ, and propose and compare several estimation methods. The proposed methods include regression estimators based on an outcome regression (OR) model, which describes the conditional distribution of the outcome given treatment and confounders, or based on a generalized PI (GPI) model, which generalizes a PI model from (2) to an arbitrary function h. An OR model implies a GPI model, so the latter is strictly more flexible. However, as we illustrate in Section 3.1, a GPI model can be difficult to estimate, and sometimes it is easier to work with an OR model. The consistency of a regression estimator, whether it is based on an OR model or a GPI model, generally depends on correct specification of the underlying model. We also consider an inverse probability weighted (IPW) estimator based on a PS model, whose consistency depends on correct specification of the PS model. Motivated by the efficient influence function for θ, we propose two doubly robust (DR), locally efficient estimators. One of them involves a PS model and an OR model, is consistent and asymptotically normal under the union of the two models, and attains the semiparametric information bound when both models are correct. The other DR estimator has the same properties with the OR model replaced by a GPI model. The second DR estimator is generally more robust, and sometimes more challenging to implement, than the first one.

In the next section, we describe these methods for observational studies and make comments on clinical trials. We then report simulation results in Section 3 and present two examples in Section 4. Concluding remarks are given in Section 5. Technical details and additional simulation results are provided in Web-Based Supplementary Materials.

2 Methodology

2.1 Overview

We restrict attention to observational studies until Section 2.7. Let T denote the actual treatment received by an individual subject. Assuming consistency or stable unit treatment value, the actual outcome is then $Y = Y(T) = TY(1) + (1-T)Y(0)$. Let $\boldsymbol{X}$ be a vector of possible confounders measured at baseline. We assume that treatment assignment is strongly ignorable[21] with respect to $\boldsymbol{X}$:

$$\mathrm{P}\{T = 1|\boldsymbol{X}, Y(0), Y(1)\} = \mathrm{P}(T = 1|\boldsymbol{X}) =: p(\boldsymbol{X}), \tag{3}$$

$$0 < p(\boldsymbol{X}) < 1 \quad \text{with probability 1.} \tag{4}$$

The observed data consist of $\boldsymbol{Z}_i = (\boldsymbol{X}_i, T_i, Y_i)$, $i = 1, \ldots, n$, which we conceptualize as independent copies of $\boldsymbol{Z} = (\boldsymbol{X}, T, Y)$.

Our goal is to use the observed data to estimate θ for a general (specified) function h. The efficient influence function in this estimation problem is given in Web Appendix A. In the rest of this section, we propose several estimators of θ, whose asymptotic representations are provided in Web Appendix B. For each estimator, an analytical variance estimate could be obtained as the sample variance of an empirical version of its influence function. This approach can be rather cumbersome to implement, depending on the specific forms of the working models involved. Alternatively, standard errors and confidence intervals could be obtained using a nonparametric bootstrap approach, which requires more computation but is easy to implement.

2.2 Regression Estimator Based on OR Model

A natural estimator of θ is motivated by the following representation:

$$\theta = \iint h(y_1, y_0) dG_1(y_1) dG_0(y_0) =: h(G_1, G_0),$$

where G_t denotes the marginal distribution of $Y(t)$, $t = 0, 1$. In general, for any measures (ν_1, ν_0), we write $h(\nu_1, y_0) = \int h(y_1, y_0) d\nu_1(y_1)$, $h(y_1, \nu_0) = \int h(y_1, y_0) d\nu_0(y_0)$, and $h(\nu_1, \nu_0) = \iint h(y_1, y_0) d\nu_1(y_1) d\nu_0(y_0)$. The ignorability assumption (3) allows G_t $(t = 0, 1)$ to be identified as

$$G_t(y) = \mathrm{P}\{Y(t) \le y\} = \mathrm{E}[\mathrm{P}\{Y(t) \le y | \boldsymbol{X}\}] = \mathrm{E}\{\mathrm{P}(Y \le y | \boldsymbol{X}, T = t)\} =: \mathrm{E}\{F_t(y|\boldsymbol{X})\},$$

The conditional distribution function $F_t(y|\boldsymbol{x})$ may be modeled as $F_t(y|\boldsymbol{x}; \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is an unknown, finite-dimensional parameter. The model $F_t(y|\boldsymbol{x}; \boldsymbol{\alpha})$ will be referred to as the OR model. Let $\widehat{\boldsymbol{\alpha}}$ be obtained by maximizing the likelihood $\prod_{i=1}^n f_{T_i}(Y_i|\boldsymbol{X}_i; \boldsymbol{\alpha})$, where $f_t(\cdot|\boldsymbol{x}; \boldsymbol{\alpha})$ is the density of $F_t(\cdot|\boldsymbol{x}; \boldsymbol{\alpha})$ with respect to a suitable measure, and let

$$\widehat{G}_t^{\mathrm{or}}(y) = n^{-1} \sum_{i=1}^n F_t(y|\boldsymbol{X}_i; \widehat{\boldsymbol{\alpha}}), \qquad t = 0, 1.$$

An OR-based regression estimator of $\theta = h(G_1, G_0)$ can then be obtained as

$$\widehat{\theta}_{\mathrm{reg}}^{\mathrm{or}} = h(\widehat{G}_1^{\mathrm{or}}, \widehat{G}_0^{\mathrm{or}}) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \iint h(y_1, y_0) F_1(dy_1|\boldsymbol{X}_i; \widehat{\boldsymbol{\alpha}}) F_0(dy_0|\boldsymbol{X}_j; \widehat{\boldsymbol{\alpha}}).$$

Under regularity conditions, $\widehat{\theta}_{\mathrm{reg}}^{\mathrm{or}}$ is consistent for θ and asymptotically normal under correct specification of the OR model. If the OR model is misspecified, then $\widehat{\theta}_{\mathrm{reg}}^{\mathrm{or}}$ may become inconsistent.

2.3 Regression Estimator Based on GPI Model

Another regression estimator, inspired by Vermeulen et al.[16], can be derived from the following expression:

$$\theta = \mathrm{E}[\mathrm{E}\{h(Y_i(1), Y_j(0))|\boldsymbol{X}_i, \boldsymbol{X}_j\}] = \mathrm{E}[\mathrm{E}\{h(Y_i, Y_j)|T_i = 1, T_j = 0, \boldsymbol{X}_i, \boldsymbol{X}_j\}] =: \mathrm{E}\{\hbar(\boldsymbol{X}_i, \boldsymbol{X}_j)\}, \tag{5}$$

where $i \neq j$ and the second equality follows from the ignorability assumption. The conditional expectation $\hbar(\boldsymbol{X}_i, \boldsymbol{X}_j)$ may be modeled as $\hbar(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is an unknown, finite-dimensional parameter. The model $\hbar(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})$ generalizes the PI model of Thas et al.[19] from (2) to an arbitrary function h, and therefore will be referred to as a GPI model. The parameter $\boldsymbol{\beta}$ can be estimated by solving estimating equations similar to those proposed by Thas et al.[19]:

$$\sum_{(i,j):T_i=1,T_j=0} \frac{\partial \hbar(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{h(Y_i, Y_j) - \hbar(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})}{V_{ij}(\boldsymbol{\beta})} = \mathbf{0}, \tag{6}$$

where $V_{ij}(\boldsymbol{\beta})$ is (an estimate of) the conditional variance of $h(Y_i, Y_j)$ given $(T_i = 1, T_j = 0, \boldsymbol{X}_i, \boldsymbol{X}_j)$ evaluated at $\boldsymbol{\beta}$. Let $\widehat{\boldsymbol{\beta}}$ denote the resulting estimate of $\boldsymbol{\beta}$; then a GPI-based regression estimator of θ is given by

$$\widehat{\theta}^{\mathrm{gpi}}_{\mathrm{reg}} = \frac{1}{n(n-1)} \sum_{i \neq j} \hbar(\boldsymbol{X}_i, \boldsymbol{X}_j; \widehat{\boldsymbol{\beta}}).$$

Under regularity conditions, $\widehat{\theta}^{\mathrm{gpi}}_{\mathrm{reg}}$ is consistent for θ and asymptotically normal under correct specification of the GPI model. If the GPI model is misspecified, then $\widehat{\theta}^{\mathrm{gpi}}_{\mathrm{reg}}$ may become inconsistent.

The estimator $\widehat{\theta}^{\mathrm{gpi}}_{\mathrm{reg}}$ is more robust than $\widehat{\theta}^{\mathrm{or}}_{\mathrm{reg}}$ because a GPI model is generally less restrictive than an OR model. Given an OR model $F_t(y|\boldsymbol{x}; \boldsymbol{\alpha})$, a GPI model can be deduced as

$$\hbar(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta}) = \iint h(y_1, y_0) F_1(dy_1|\boldsymbol{X}_i; \boldsymbol{\alpha}) F_0(dy_0|\boldsymbol{X}_j; \boldsymbol{\alpha}),$$

where $\boldsymbol{\beta}$ is, in general, a vector-valued function of $\boldsymbol{\alpha}$. However, there may be practical reasons to prefer $\widehat{\theta}^{\text{or}}_{\text{reg}}$ over $\widehat{\theta}^{\text{gpi}}_{\text{reg}}$ in some situations. First, as we demonstrate in Section 3.1, a GPI model deduced from a given OR model can be quite complicated, and equation (6) may be difficult to solve. Techniques for specifying and estimating an OR model are much better developed in comparison. Second, the computational complexity for fitting a GPI model is clearly higher than that for fitting an OR model; this is a real issue in analyzing the example in Section 4.1. Third, when the underlying OR model is correct, $\widehat{\theta}^{\text{or}}_{\text{reg}}$ can be expected to be more efficient than $\widehat{\theta}^{\text{gpi}}_{\text{reg}}$.

2.4 IPW Estimator Based on PS Model

An alternative approach to estimating θ is based on the observation that, for two independent subjects ($i \neq j$),

$$\theta = \mathrm{E}\left[\frac{T_i(1-T_j)h(Y_i,Y_j)}{p(\boldsymbol{X}_i)\{1-p(\boldsymbol{X}_j)\}}\right], \tag{7}$$

which follows from standard conditioning arguments. Let the PS $p(\boldsymbol{x})$ be modeled as $p(\boldsymbol{x};\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is an unknown, finite-dimensional parameter. The model $p(\boldsymbol{x};\boldsymbol{\gamma})$ will be referred to as the PS model. Let $\widehat{\boldsymbol{\gamma}}$ be obtained by maximizing the likelihood $\prod_{i=1}^n p(\boldsymbol{X}_i;\boldsymbol{\gamma})^{T_i}\{1-p(\boldsymbol{X}_i;\boldsymbol{\gamma})\}^{1-T_i}$. Then we can estimate θ with the IPW estimator

$$\widehat{\theta}_{\text{ipw}} = \sum_{i=1}^n\sum_{j=1}^n \left[\frac{T_i(1-T_j)h(Y_i,Y_j)}{p(\boldsymbol{X}_i;\widehat{\boldsymbol{\gamma}})\{1-p(\boldsymbol{X}_j;\widehat{\boldsymbol{\gamma}})\}}\right] \Bigg/ \sum_{i=1}^n\sum_{j=1}^n \left[\frac{T_i(1-T_j)}{p(\boldsymbol{X}_i;\widehat{\boldsymbol{\gamma}})\{1-p(\boldsymbol{X}_j;\widehat{\boldsymbol{\gamma}})\}}\right].$$

It is easy to see that $\widehat{\theta}_{\text{ipw}} = h(\widehat{G}^{\text{ipw}}_1, \widehat{G}^{\text{ipw}}_0)$, where

$$\widehat{G}^{\text{ipw}}_t(y) = \sum_{i=1}^n \frac{I(T_i=t)I(Y_i\le y)}{p(\boldsymbol{X}_i;\widehat{\boldsymbol{\gamma}})^t\{1-p(\boldsymbol{X}_i;\widehat{\boldsymbol{\gamma}})\}^{1-t}} \Bigg/ \sum_{i=1}^n \frac{I(T_i=t)}{p(\boldsymbol{X}_i;\widehat{\boldsymbol{\gamma}})^t\{1-p(\boldsymbol{X}_i;\widehat{\boldsymbol{\gamma}})\}^{1-t}}.$$

Under regularity conditions, $\widehat{\theta}_{\text{ipw}}$ is consistent for θ and asymptotically normal if the PS model is correct. If the PS model is misspecified, $\widehat{\theta}_{\text{ipw}}$ is generally inconsistent.

2.5 DR Estimator Based on PS and OR Models

Each of the methods considered so far relies on a different model for consistent estimation of θ. More robust methods are desirable because in practice it can be difficult to specify an (approximately) correct parametric model. There is an extensive literature on doubly robust estimation in causal inference[2,3,4,5]. Such estimators are frequently motivated by semiparametric theory[22,4]. In the present context, the parameter θ, as a functional of the joint distribution of $\boldsymbol{Z} = (\boldsymbol{X}, T, Y)$, is pathwise differentiable with efficient influence function

$$\phi_{\text{eff}}(\boldsymbol{Z}) = \frac{T\{h(Y, G_0) - \theta\}}{p(\boldsymbol{X})} + \frac{(1-T)\{h(G_1, Y) - \theta\}}{1 - p(\boldsymbol{X})} + \{T - p(\boldsymbol{X})\}\left\{\frac{h(G_1, F_0(\cdot|\boldsymbol{X})) - \theta}{1 - p(\boldsymbol{X})} - \frac{h(F_1(\cdot|\boldsymbol{X}), G_0) - \theta}{p(\boldsymbol{X})}\right\}$$

in a nonparametric model that leaves $F_t(y|\boldsymbol{x})$ and $p(\boldsymbol{x})$ unspecified. Motivated by this result, we consider estimators of the form

$$\frac{1}{2n}\sum_{i=1}^{n}\left[\frac{T_i h(Y_i, \widehat{G}_0)}{p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})} + \frac{(1-T_i)h(\widehat{G}_1, Y_i)}{1 - p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})} + \{T_i - p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})\}\left\{\frac{h(\widehat{G}_1, F_0(\cdot|\boldsymbol{X}_i; \widehat{\boldsymbol{\alpha}}))}{1 - p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})} - \frac{h(F_1(\cdot|\boldsymbol{X}_i; \widehat{\boldsymbol{\alpha}}), \widehat{G}_0)}{p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})}\right\}\right], \quad (8)$$

where $\widehat{G}_t$ $(t = 0, 1)$ is a generic estimator of G_t. Possible choices for $\widehat{G}_t$ include $\widehat{G}_t^{\text{or}}$ and $\widehat{G}_t^{\text{ipw}}$. The estimator $\widehat{G}_t^{\text{or}}$ is based on the OR model, and its substitution into (8) produces an estimator of θ that is consistent for θ under the OR model but not necessarily under the PS model. The estimator $\widehat{G}_t^{\text{ipw}}$ is based on the PS model, and the resulting estimator of θ is consistent under the PS model but not necessarily under the OR model. To achieve double robustness, we propose to substitute into (8) the following DR estimator:

$$\widehat{G}_t^{\text{dr}}(y) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I(T_i = t)I(Y_i \le y)}{p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})^t\{1 - p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})\}^{1-t}} - \frac{I(T_i = t) - p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})^t\{1 - p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})\}^{1-t}}{p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})^t\{1 - p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})\}^{1-t}} F_t(y|\boldsymbol{X}_i; \widehat{\boldsymbol{\alpha}})\right].$$

The resulting estimator of θ will be denoted by $\widehat{\theta}_{\mathrm{dr1}}$. It can be shown that $\widehat{\theta}_{\mathrm{dr1}} = h(\widehat{G}_1^{\mathrm{dr}}, \widehat{G}_0^{\mathrm{dr}})$, which implies that $\widehat{\theta}_{\mathrm{dr1}}$ is consistent under the union of the PS and OR models. Furthermore, under the same union model, $\widehat{\theta}_{\mathrm{dr1}}$ is asymptotically linear with influence function

$$\begin{aligned}
\phi_{\mathrm{dr1}}(\boldsymbol{Z}) = {} & \frac{T\{h(Y, G_0) - \theta\}}{p(\boldsymbol{X}; \boldsymbol{\gamma}^*)} - \frac{\{T - p(\boldsymbol{X}; \boldsymbol{\gamma}^*)\}\{h(F_1(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*), G_0) - \theta\}}{p(\boldsymbol{X}; \boldsymbol{\gamma}^*)} \\
& - \frac{\partial}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \mathrm{E}\left[\frac{\{T - p(\boldsymbol{X}; \boldsymbol{\gamma}^*)\} h(F_1(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*), G_0)}{p(\boldsymbol{X}; \boldsymbol{\gamma}^*)}\right] \boldsymbol{\psi}_{\widehat{\boldsymbol{\alpha}}}(\boldsymbol{Z}) \\
& + \frac{\partial}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \mathrm{E}\left[\frac{T\{h(Y, G_0) - h(F_1(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*), G_0)\}}{p(\boldsymbol{X}; \boldsymbol{\gamma}^*)}\right] \boldsymbol{\psi}_{\widehat{\boldsymbol{\gamma}}}(\boldsymbol{Z}) \\
& + \frac{(1 - T)\{h(G_1, Y) - \theta\}}{1 - p(\boldsymbol{X}; \boldsymbol{\gamma}^*)} + \frac{\{T - p(\boldsymbol{X}; \boldsymbol{\gamma}^*)\}\{h(G_1, F_0(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*)) - \theta\}}{1 - p(\boldsymbol{X}; \boldsymbol{\gamma}^*)} \\
& + \frac{\partial}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \mathrm{E}\left[\frac{\{T - p(\boldsymbol{X}; \boldsymbol{\gamma}^*)\} h(G_1, F_0(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*))}{1 - p(\boldsymbol{X}; \boldsymbol{\gamma}^*)}\right] \boldsymbol{\psi}_{\widehat{\boldsymbol{\alpha}}}(\boldsymbol{Z}) \\
& - \frac{\partial}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \mathrm{E}\left[\frac{(1 - T)\{h(G_1, Y) - h(G_1, F_0(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*))\}}{1 - p(\boldsymbol{X}; \boldsymbol{\gamma}^*)}\right] \boldsymbol{\psi}_{\widehat{\boldsymbol{\gamma}}}(\boldsymbol{Z}),
\end{aligned}$$

where $\boldsymbol{\alpha}^*$ and $\boldsymbol{\gamma}^*$ are the respective probability limits of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\gamma}}$, and $\boldsymbol{\psi}_{\widehat{\boldsymbol{\alpha}}}(\boldsymbol{Z})$ and $\boldsymbol{\psi}_{\widehat{\boldsymbol{\gamma}}}(\boldsymbol{Z})$ are the respective influence functions for $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\gamma}}$. If the OR model is correctly specified, then $\boldsymbol{\alpha}^*$ is the true value of $\boldsymbol{\alpha}$ and the terms involving $\boldsymbol{\psi}_{\widehat{\boldsymbol{\gamma}}}(\boldsymbol{Z})$ vanish. Similarly, if the PS model is correctly specified, then $\boldsymbol{\gamma}^*$ is the true value of $\boldsymbol{\gamma}$ and the terms involving $\boldsymbol{\psi}_{\widehat{\boldsymbol{\alpha}}}(\boldsymbol{Z})$ vanish. If both models are correct, then $\phi_{\mathrm{dr1}}(\boldsymbol{Z}) = \phi_{\mathrm{eff}}(\boldsymbol{Z})$ and $\widehat{\theta}_{\mathrm{dr1}}$ attains the semiparametric information bound. In a finite sample, $\widehat{\theta}_{\mathrm{dr1}}$ could take values outside the range of h, in which case we truncate $\widehat{\theta}_{\mathrm{dr1}}$ into the range of h. This slight modification does not alter the consistency or asymptotic normality of $\widehat{\theta}_{\mathrm{dr1}}$.

2.6 DR Estimator Based on PS and GPI Models

Finally, consider the estimator

$$\begin{aligned}
\widehat{\theta}_{\mathrm{dr2}} = \frac{1}{n(n-1)} \sum_{i \neq j} \Bigg[& \frac{T_i(1 - T_j) h(Y_i, Y_j)}{p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})\{1 - p(\boldsymbol{X}_j; \widehat{\boldsymbol{\gamma}})\}} \\
& - \frac{T_i(1 - T_j) - p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})\{1 - p(\boldsymbol{X}_j; \widehat{\boldsymbol{\gamma}})\}}{p(\boldsymbol{X}_i; \widehat{\boldsymbol{\gamma}})\{1 - p(\boldsymbol{X}_j; \widehat{\boldsymbol{\gamma}})\}} \hbar(\boldsymbol{X}_i, \boldsymbol{X}_j; \widehat{\boldsymbol{\beta}}) \Bigg],
\end{aligned}$$

which generalizes the augmented estimator of Vermeulen et al. [16] for randomized experiments. That $\widehat{\theta}_{\mathrm{dr2}}$ is DR, as the subscript suggests, can be seen as follows. Under regularity conditions, we can expect $\widehat{\theta}_{\mathrm{dr2}}$ to converge in probability to

$$\mathrm{E}\left[\frac{T_i(1-T_j)h(Y_i,Y_j)}{p(\boldsymbol{X}_i;\boldsymbol{\gamma}^*)\{1-p(\boldsymbol{X}_j;\boldsymbol{\gamma}^*)\}} - \frac{T_i(1-T_j)-p(\boldsymbol{X}_i;\boldsymbol{\gamma}^*)\{1-p(\boldsymbol{X}_j;\boldsymbol{\gamma}^*)\}}{p(\boldsymbol{X}_i;\boldsymbol{\gamma}^*)\{1-p(\boldsymbol{X}_j;\boldsymbol{\gamma}^*)\}}\hbar(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta}^*)\right] \tag{9}$$

$$= \mathrm{E}\left[\hbar(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta}^*) + \frac{T_i(1-T_j)\{h(Y_i,Y_j)-\hbar(\boldsymbol{X}_i,\boldsymbol{X}_j;\boldsymbol{\beta}^*)\}}{p(\boldsymbol{X}_i;\boldsymbol{\gamma}^*)\{1-p(\boldsymbol{X}_j;\boldsymbol{\gamma}^*)\}}\right], \tag{10}$$

where $\boldsymbol{\beta}^*$ denotes the probability limit of $\widehat{\boldsymbol{\beta}}$ and the subscripts i and j denote two independent subjects. If the PS model is correct, then $\boldsymbol{\gamma}^*$ equals the true value of $\boldsymbol{\gamma}$ and it is easy to see that (9) becomes (7). If the GPI model is correct, then $\boldsymbol{\beta}^*$ is the true value of $\boldsymbol{\beta}$ and (10) is equivalent to (5). Hence $\widehat{\theta}_{\mathrm{dr2}}$ is consistent for θ under the union of the PS and GPI models. Furthermore, under the same union model, $\widehat{\theta}_{\mathrm{dr2}}$ is asymptotically linear with influence function

$$\phi_{\mathrm{dr2}}(\boldsymbol{Z}) = 2k'_{\mathrm{dr2}}(\boldsymbol{Z};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*) - 2\theta + \frac{\partial\,\mathrm{E}\{k'_{\mathrm{dr2}}(\boldsymbol{Z};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*)\}}{\partial\boldsymbol{\beta}}\boldsymbol{\psi}_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{Z}) + \frac{\partial\,\mathrm{E}\{k'_{\mathrm{dr2}}(\boldsymbol{Z};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*)\}}{\partial\boldsymbol{\gamma}}\boldsymbol{\psi}_{\widehat{\boldsymbol{\gamma}}}(\boldsymbol{Z}),$$

where $\boldsymbol{\psi}_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{Z})$ is the influence function for $\widehat{\boldsymbol{\beta}}$,

$$\begin{aligned} k'_{\mathrm{dr2}}(\boldsymbol{Z};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*) = {} & \big[\mathrm{E}_{\boldsymbol{X}'}\{\hbar(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}^*)\} + \mathrm{E}_{\boldsymbol{X}'}\{\hbar(\boldsymbol{X}',\boldsymbol{X};\boldsymbol{\beta}^*)\}\big]/2 \\ & + \frac{1}{2}\,\mathrm{E}_{\boldsymbol{X}'}\left[\frac{T\{1-p(\boldsymbol{X}')\}\{h(Y,F_0(\cdot|\boldsymbol{X}'))-\hbar(\boldsymbol{X},\boldsymbol{X}';\boldsymbol{\beta}^*)\}}{p(\boldsymbol{X};\boldsymbol{\gamma}^*)\{1-p(\boldsymbol{X}';\boldsymbol{\gamma}^*)\}}\right] \\ & + \frac{1}{2}\,\mathrm{E}_{\boldsymbol{X}'}\left[\frac{(1-T)p(\boldsymbol{X}')\{h(F_1(\cdot|\boldsymbol{X}'),Y)-\hbar(\boldsymbol{X}',\boldsymbol{X};\boldsymbol{\beta}^*)\}}{p(\boldsymbol{X}';\boldsymbol{\gamma}^*)\{1-p(\boldsymbol{X};\boldsymbol{\gamma}^*)\}}\right], \end{aligned}$$

$\boldsymbol{X}' \sim \boldsymbol{X}$ independently of $\boldsymbol{Z}$, and $\mathrm{E}_{\boldsymbol{X}'}$ denotes expectation with respect to $\boldsymbol{X}'$. It is easy to see that $\partial\,\mathrm{E}\{k'_{\mathrm{dr2}}(\boldsymbol{Z};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*)\}/\partial\boldsymbol{\beta} = \mathbf{0}$ under the PS model and that $\partial\,\mathrm{E}\{k'_{\mathrm{dr2}}(\boldsymbol{Z};\boldsymbol{\beta}^*,\boldsymbol{\gamma}^*)\}/\partial\boldsymbol{\gamma} = \mathbf{0}$ under the GPI model. If both models are correct, then $\phi_{\mathrm{dr2}}(\boldsymbol{Z}) = 2k'_{\mathrm{dr2}}(\boldsymbol{Z};\boldsymbol{\beta},\boldsymbol{\gamma}) - 2\theta = \phi_{\mathrm{eff}}(\boldsymbol{Z})$ and $\widehat{\theta}_{\mathrm{dr2}}$ attains the semiparametric information bound. Like $\widehat{\theta}_{\mathrm{dr1}}$, $\widehat{\theta}_{\mathrm{dr2}}$ can take values outside the range of h and will be truncated in our implementation. Compared with

$\widehat{\theta}_{\mathrm{dr1}}$, $\widehat{\theta}_{\mathrm{dr2}}$ has some theoretical advantages in that its double robustness and local efficency are more broadly applicable. On the other hand, $\widehat{\theta}_{\mathrm{dr1}}$ may be easier to use in practice.

2.7 Remarks on Clinical Trials

With h defined by (2), estimation of θ in a randomized clinical trial has been studied thoroughly by Vermeulen et al.[16]. Here we comment briefly on the generalization of their results to an arbitrary function h as well as the use of an OR model in this setting.

By design, T is independent of all baseline variables including $\boldsymbol{X}$ and the potential outcomes, which allows θ to be estimated empirically by

$$\widehat{\theta}_{\mathrm{emp}} = \frac{1}{n_0 n_1} \sum_{i=1}^{n} \sum_{j=1}^{n} T_i(1-T_j)h(Y_i, Y_j),$$

where $n_t = \sum_{i=1}^{n} I(T_i = t)$, $t = 0, 1$. Using the theory of U-statistics[23], it can be shown that $\widehat{\theta}_{\mathrm{emp}}$ is consistent for θ and asymptotically linear with influence function

$$\phi_{\mathrm{emp}}(Z) = \frac{T\{h(Y, G_0) - \theta\}}{\pi} + \frac{(1-T)\{h(G_1, Y) - \theta\}}{1-\pi},$$

where $\pi = \mathrm{P}(T = 1)$. It is easy to see that $\widehat{\theta}_{\mathrm{emp}}$ is efficient among nonparametric estimators of θ based on $\{(T_i, Y_i), i = 1, \ldots, n\}$. However, $\widehat{\theta}_{\mathrm{emp}}$ may be inefficient in the present context because it does not utilize the available information in $\boldsymbol{X}$.

It can be argued as in Tsiatis[4] and Vermeulen et al.[16] that any nonparametric estimator of θ that is regular and asymptotically linear has influence function $\phi_{\mathrm{emp}}(\boldsymbol{Z}) + a(T, \boldsymbol{X})$, where a is such that $\mathrm{E}\{a(T, \boldsymbol{X})|\boldsymbol{X}\} = 0$ and $\mathrm{var}\{a(T, \boldsymbol{X})\} < \infty$. The asymptotic variance of the estimator is minimized by taking $a(T, \boldsymbol{X})$ to be

$$a_{\mathrm{opt}}(T, \boldsymbol{X}) = (T - \pi)\left\{\frac{h(G_1, F_0(\cdot|\boldsymbol{X})) - \theta}{1-\pi} - \frac{h(F_1(\cdot|\boldsymbol{X}), G_0) - \theta}{\pi}\right\}.$$

To take advantage of this result, we can use the augmented estimator of Vermeulen et al.[16],

which in the present setting can be written as

$$\widehat{\theta}^{\text{gpi}}_{\text{aug}} = \widehat{\theta}_{\text{emp}} + \sum_{i \neq j} \left\{ \frac{1}{n(n-1)} - \frac{T_i(1-T_j)}{n_0 n_1} \right\} \hbar(\boldsymbol{X}_i, \boldsymbol{X}_j; \widehat{\boldsymbol{\beta}}),$$

where $\hbar(\boldsymbol{X}_i, \boldsymbol{X}_j; \widehat{\boldsymbol{\beta}})$ is the same estimated GPI model described earlier. If the GPI model is correct, $\widehat{\theta}^{\text{gpi}}_{\text{aug}}$ has influence function equal to $\phi_{\text{emp}}(\boldsymbol{Z}) + a_{\text{opt}}(T, \boldsymbol{X})$ and therefore attains the minimum variance. If the GPI model is incorrect, $\widehat{\theta}^{\text{gpi}}_{\text{aug}}$ remains consistent for θ and asymptotically normal.

Alternatively, we could obtain an augmented estimator based on an OR model, which is also locally efficient (in a narrower sense) and may be easier to use in some situations. The OR-based augmented estimator is

$$\widehat{\theta}^{\text{or}}_{\text{aug}} = \widehat{\theta}_{\text{emp}} + \frac{1}{n} \sum_{i=1}^{n} \left[(T_i - \widehat{\pi}) \left\{ \frac{h(\widehat{G}^{\text{emp}}_1, F_0(\cdot|\boldsymbol{X}_i; \widehat{\boldsymbol{\alpha}}))}{1-\widehat{\pi}} - \frac{h(F_1(\cdot|\boldsymbol{X}_i; \widehat{\boldsymbol{\alpha}}), \widehat{G}^{\text{emp}}_0)}{\widehat{\pi}} \right\} \right],$$

where $\widehat{\pi} = n_1/n$ and $\widehat{G}^{\text{emp}}_t(y) = n_t^{-1} \sum_{i=1}^{n} I(T_i = t, Y_i \leq y)$, $t = 0, 1$. Under standard regularity conditions, $\widehat{\theta}^{\text{or}}_{\text{aug}}$ is consistent for θ and asymptotically linear with influence function

$$\phi^{\text{or}}_{\text{aug}}(Z) = \phi_{\text{emp}}(Z) + (T - \pi) \left\{ \frac{h(G_1, F_0(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*)) - \theta_0^*}{1-\pi} - \frac{h(F_1(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*), G_0) - \theta_1^*}{\pi} \right\},$$

where $\theta_0^* = \text{E}\{h(G_1, F_0(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*))\}$ and $\theta_1^* = \text{E}\{h(F_1(\cdot|\boldsymbol{X}; \boldsymbol{\alpha}^*), G_0)\}$. If the OR model is correct, then $\boldsymbol{\alpha}^*$ is the true value of $\boldsymbol{\alpha}$, $\theta_0^* = \theta_1^* = \theta$, $\phi^{\text{or}}_{\text{aug}}(\boldsymbol{Z}) = \phi_{\text{emp}}(\boldsymbol{Z}) + a_{\text{opt}}(\boldsymbol{Z})$, and hence $\widehat{\theta}^{\text{or}}_{\text{aug}}$ attains the minimum variance. If the OR model is incorrect, $\widehat{\theta}^{\text{or}}_{\text{aug}}$ remains consistent for θ and asymptotically normal.

3 Simulation

We now evaluate and compare the finite-sample performance of the various methods in simulation experiments, with h defined by (1). Each experiment is based on 1000 replicate samples, each of which consists of 500 independent copies of $(\boldsymbol{X}, T, Y)$. The covariate vector

$\boldsymbol{X}$ has three components (X_1, X_2, X_3), which are independent and identically distributed as standard normal. Given $\boldsymbol{X}$, T follows a logistic regression model with

$$\text{logit}\{\text{P}(T = 1|\boldsymbol{X})\} = (1, X_1, X_2, X_3, X_1X_2, X_1X_3)\boldsymbol{\gamma}, \tag{11}$$

where $\boldsymbol{\gamma}$ is either $\mathbf{0}$ (for a randomized clinical trial) or $(0, -1, 1, -1, 1, -1)^{\text{T}}$ (for an observational study). The outcome variable Y may be discrete or continuous.

3.1 Discrete Outcome

In this subsection, Y is a discrete ordinal outcome with three levels $(0, 1, 2)$. Given $(\boldsymbol{X}, T)$, Y is generated from a multinomial logistic regression model with

$$\log\left\{\frac{\text{P}(Y = j|\boldsymbol{X}, T)}{\text{P}(Y = 0|\boldsymbol{X}, T)}\right\} = (1, X_1, X_2, X_3, T, TX_2, TX_3)\boldsymbol{\alpha}_j \quad (j = 1, 2), \tag{12}$$

where $\boldsymbol{\alpha}_1 = (0, 1, 0.5, 1, 0 \text{ or } 1, -1, -1)^{\text{T}}$ and $\boldsymbol{\alpha}_2 = 2\boldsymbol{\alpha}_1$. The true value of θ is virtually zero or positive according as the regression coefficient for T in $\boldsymbol{\alpha}_1$ equals 0 or 1.

We first consider the case of randomized clinical trials ($\boldsymbol{\gamma} = \mathbf{0}$). For each simulated sample, we calculate the empirical estimator and the augmented estimators with (approximately) correct and incorrect OR and GPI models. The correct OR model is given by (12), and the incorrect model takes the same form with X_1 replaced by $I(X_1 > 0)$. The correct GPI model deduced from (12) is very complicated, highly non-linear, and not estimable with the `pim` package in R. Although there are packages for fitting non-linear models, they tend to be unstable and often fail to converge. For practicality, we therefore consider the following logistic model:

$$\begin{aligned}\text{logit}\,\text{E}\{I(Y_i > Y_j) + 0.5I(Y_i = Y_j)|T_i = 1, T_j = 0, \boldsymbol{X}_i, \boldsymbol{X}_j\} \\ = (1, X_{1i} - X_{1j}, X_{2i} - X_{2j}, X_{3i} - X_{3j}, X_{2i}, X_{3i})\boldsymbol{\beta},\end{aligned}$$

which is considered "approximately correct" in that it has the "right" linear terms. We also consider a (more) incorrect GPI model that takes the same form with X_1 replaced by $I(X_1 > 0)$. The resulting five estimators are compared in the upper portion of Table 1 (under "discrete outcome") in terms of empirical bias and standard deviation. Table 1 shows that all three estimators are nearly unbiased but they differ substantially in precision. As expected, the best precision is attained by the augmented estimator based on a correct OR model. The augmented estimator based on an approximately correct GPI model is similarly efficient when $\theta \approx 0$ and slightly less efficient when $\theta > 0$. The augmented estimators become less efficient when the working models are misspecified, but still more efficient than the empirical estimator. To assess the improvement in efficiency, we calculate the relative efficiency of an augmented estimator as the inverse ratio of its (estimated) variance to that of the empirical estimator. In Table 1, the relative efficiency ranges from 1.26 to 1.47 for a discrete outcome.

Next, we consider the case of observational studies ($\boldsymbol{\gamma} \neq \mathbf{0}$). For each simulated sample, we calculate the regression, IPW and DR estimators with (approximately) correct and incorrect OR, GPI and PS models. The (approximately) correct and incorrect OR and GPI models are as described in the last paragraph. The correct PS model is given by (11), and the incorrect PS model takes the same form with X_1 replaced by $I(X_1 > 0)$. The estimators are compared in the upper portion of Table 2 (under "discrete outcome") in terms of empirical bias, standard deviation and root mean squared error. As expected, the regression estimators become biased when the OR or GPI model is misspecified, as does the IPW estimator when the PS model is misspecified. The DR estimators remain nearly unbiased under the union of the OR/GPI and PS models. With correct working models, the regression estimators are usually the most efficient, and the IPW estimator tends to be the least efficient. With misspecified models, the regression estimators still have the smallest standard deviations,

but the DR estimators can be more variable than the IPW estimator. In terms of mean squared error, the regression estimators appear preferable in this particular setting. It seems inconsequential whether to use an OR model or a GPI model in constructing a regression or DR estimator.

3.2 Continuous Outcome

In this subsection, Y is a continuous outcome following a skewed distribution. Given $(\boldsymbol{X}, T)$, Y is generated from a Box-Cox-transformed linear model[24] with

$$F_t(y|\boldsymbol{x}) = \Phi\left(\frac{b(y;\lambda) - (1, x_1, x_2, x_3, t, tx_2, tx_3)\boldsymbol{\eta}}{\sigma}\right), \tag{13}$$

where $b(y;\lambda)$ is the Box-Cox transformation with parameter λ: $\log y$ if $\lambda = 0$ or $(y^\lambda - 1)/\lambda$ if $\lambda \neq 0$. For data generation, we set $\lambda = 0.2$, $\sigma = 1$ and $\boldsymbol{\eta} = (10, 1, 0.5, 1, 0 \text{ or } 1, -1, -1)^{\mathrm{T}}$. The true value of θ is zero or positive according as the regression coefficient for T in $\boldsymbol{\eta}$ equals 0 or 1.

In the case of randomized clinical trials ($\boldsymbol{\gamma} = \mathbf{0}$), we compare the empirical estimator and the augmented estimators with correct and incorrect OR and GPI models. The correct OR model is given by (13), and the incorrect OR model takes the same form with X_1 replaced by $I(X_1 > 0)$. The correct GPI model is the following probit regression model:

$$\mathrm{P}(Y_i > Y_j | T_i = 1, T_j = 0, \boldsymbol{X}_i, \boldsymbol{X}_j) = \Phi\big((1, X_{1i} - X_{1j}, X_{2i} - X_{2j}, X_{3i} - X_{3j}, X_{2i}, X_{3i})\boldsymbol{\beta}\big),$$

where Φ is the standard normal distribution function, and the incorrect GPI model takes the same form with X_1 replaced by $I(X_1 > 0)$. The results, shown in the lower portion of Table 1 (under "continuous outcome"), are qualitatively consistent with the analogous results for a discrete outcome. The relative efficiency of the augmented estimator ranges from 1.39 to 1.86 in the present setting.

In the case of observational studies ($\boldsymbol{\gamma} \neq \mathbf{0}$), we compare the regression, IPW and DR estimators with correct and incorrect OR, GPI and PS models. The correct and incorrect OR and GPI models are as described in the last paragraph. The correct and incorrect PS models are the same as those for a discrete outcome. The results, shown in the lower portion of Table 2 (under "continuous outcome"), are largely consistent with the analogous results for a discrete outcome.

In Web Appendix C, we describe another simulation experiment in which the OR model is incorrect but the deduced GPI model is correct, which helps illustrate the extra flexibility of the GPI model. The results are similar to those in Tables 1 and 2 for a continuous outcome, and therefore not reported here.

4 Examples

4.1 The AFFIRM Study

The Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) study is a randomized clinical trial comparing two long-term treatment strategies for atrial fibrillation (AF) in patients who had a high risk of stroke or death[25]. The study enrolled 4060 patients, who were randomized $1:1$ to either a rhythm-control treatment strategy ($t = 1$) consisting of cardioversion and treatment with antiarrhythmic drugs to maintain sinus rhythm, or a rate-control treatment strategy ($t = 0$) that allowed AF to persist while controlling the ventricular response to AF. In the original analysis plan, the primary outcome was death and an important secondary outcome was a composite of death with serious adverse events (SAEs; disabling stroke, disabling anoxic encephalopathy, major bleeding, and cardiac arrest). To illustrate the proposed methodology, we consider here a three-level ordinal outcome based on death and SAEs at two years of treatment. Specifically, $Y = 0$ if a patient died within

two years, 1 if a patient survived two years with one or more SAEs, and 2 if a patient survived two years with no SAEs. A total of 102 (2.5%) patients were lost to follow-up within two years without a death report, and their outcome status cannot be determined with the available information. Their outcome data are imputed under the assumption that no death or SAEs occurred between the last evaluation and two years (because such events are usually reported). Although this assumption may not be correct for every patient, its possible violation is likely to have a small impact given the small percentage of the affected patients and their largely even distribution in the two treatment groups (50 rhythm-control; 52 rate-control). The imputed outcome data are summarized in Table 3.

Based on the data in Table 3, the empirical estimate of θ, with h defined by (1), is -0.017, with a nonparametric bootstrap standard error of 0.011 (based on 1000 bootstrap samples). To obtain an augmented estimate, we work with the following baseline covariates: age, gender, minority race (caucasian or not), history of congestive heart failure (yes or no), duration of qualifying AF ($<$ or ≥ 2 days), first episode of AF (as opposed to recurrent AF), and previous failure of an antiarrhythmic drug. Our OR model is a multinomial logistic regression model like (12) that includes treatment, baseline covariates as well as interactions between treatment and all covariates. The OR-based augmented estimate of θ is obtained as -0.020, with a nonparametric bootstrap standard error of 0.010 (based on 1000 bootstrap samples). A prohibitive amount of computation would be required to fit a similar GPI model to these data. Considering the simulation results in Section 3, we have therefore omitted the GPI-based augmented estimate in this analysis. Based on the available results, a Wald test of $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$ at level 0.05 would be marginally non-significant under the empirical approach and marginally significant under the augmented approach.

4.2 The Consortium on Safe Labor

The Consortium on Safe Labor (CSL) is a large observational study designed to describe contemporary labor progression in the United States[26]. Our causal question here pertains to the effect of epidural analgesia on the duration of the second stage of labor (from full cervical dilation to delivery of the fetus). The outcome apparently follows a skewed distribution, which casts doubts on the appropriateness of the average treatment effect. Zhang et al.[6] use quantiles to characterize the causal effect of epidural anagesia; however, their approach depends on the choice of quantiles and does not produce an overall effect measure. Here we use θ, with h defined by (1), to quantify the overall effect of epidural anagesia on duration of second-stage labor.

Our causal analysis is based on 3,838 women in the CSL who delivered at week 40 with complete information on the outcome (second-stage labor duration in minutes), the treatment (epidural), and all relevant covariates (identified prospectively as maternal age, maternal body mass index, birth weight and whether induction of labor was performed). The PS model is a standard logistic regression model with the aforementioned covariates as the linear terms (in addition to an intercept). The OR model is a Box-Cox-transformed linear model like (13) with $\boldsymbol{\alpha} = (\lambda, \boldsymbol{\eta}, \sigma^2)$ and with interactions between treatment and all covariates. The GPI model is a probit regression model deduced from the OR model.

Table 4 shows the point estimates of θ obtained using the regression, IPW and DR methods, together with nonparametric bootstrap standard errors (based on 1,000 bootstrap samples). The five point estimates are very similar to each other and, considering the small standard errors, significantly greater than 0. Assuming that treatment assignment is strongly ignorable and that at least one of the OR, GPI and PS models is nearly correct, the results in Table 3 clearly indicate that epidural prolongs second-stage labor, in the sense that a second-stage labor with epidural is at least 10% more likely to be longer than (as opposed

to shorter than) one without epidural for two randomly chosen women.

5 Discussion

Mann–Whitney-type effect measures have been advocated for clinical trials because of their clinical relevance and interpretability, but their estimation in observational studies has not received much attention in the statistics literature. This article provides new results and methods for estimating such causal effects in observational studies and clinical trials. For observational studies, the regression and DR estimators appear more competitive than the IPW estimator, and their suitability to a given application will depend on the available information and the relative importance of bias versus efficiency. If one is primarily concerned about bias, then the DR estimator may be preferable. The regression estimator is more efficient when the OR/GPI model is correctly specified. For clinical trials, we have generalized the augmented estimator of Vermeulen et al.[16] to an arbitrary function h and have proposed a new augmented estimator based on an OR model, which may be easier to use in some situations.

Although we have focused on estimation in this article, the robustness and efficiency results have immediate implications on the validity and power of Wald tests of statistical hypotheses about θ. Thus, the proposed methods can still be useful if one is primarily interested in hypothesis testing as opposed to estimating θ.

The proposed methods are based on parametric OR, GPI and PS models. It will be of interest to replace these working models with nonparametric and semiparametric models, which would add a great deal of robustness to the methods. Another area of future research is how to incorporate variable selection techniques such as the lasso into the OR and PS models when the dimension of X is very high (possibly greater than n).

Acknowledgements

Liu's research was supported by the Hong Kong Polytechnic University (grant # G-YBCU) and the National Natural Science Foundation of China (grant # 11401502). A part of this research was conducted while the first author was visiting the Hong Kong Polytechnic University. This manuscript was prepared using AFFIRM Research Materials obtained from the National Heart, Lung and Blood Institute (NHLBI) Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the AFFIRM Team or the NHLBI.

References

[1] Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

[2] van der Laan MJ, Robins JM (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Spring-Verlag, New York.

[3] Bang H, Robins JM (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–972.

[4] Tsiatis AA (2006). *Semiparametric Theory and Missing Data.* Springer, New York.

[5] van der Laan MJ, Rose S (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York.

[6] Zhang Z, Chen Z, Troendle JF, Zhang J (2012). Causal inference on quantiles with an obstetric application. *Biometrics*, 68, 697–706.

[7] Agresti A (2013). *Categorical Data Analysis*, 3rd ed. John Wiley and Sons, Hoboken, NJ.

[8] Wilcoxon F (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80-83.

[9] Mann HB, Whitney DR (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50–60.

[10] Acion L, Peterson JJ, Temple S, Arndt S (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, 25, 591–602.

[11] Brumback LC, Pepe MS, Alonzo TA (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine*, 25, 575–590.

[12] Newcombe RG (2006). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. *Statistics in Medicine*, 25, 543–557.

[13] Shih WJ, Quan H (1997). Testing for treatment differences with dropouts present in clinical trials: a composite approach. *Statistics in Medicine*, 16, 1225–1239.

[14] Liu W, Zhang Z, Nie L, Soon G (2017). A case study in personalized medicine: rilpivirine versus efavirenz for treatment-naive HIV patients. *Journal of the American Statistical Association*, in press (DOI: 10.1080/01621459.2017.1280404).

[15] Wang C, Scharfstein DO, Colantuoni E, Girard TD, Yan Y (2017). Inference in randomized trials with death and missingness. *Biometrics*, in press (DOI: 10.1111/biom.12594).

[16] Vermeulen K, Thas O, Vansteelandt S (2015). Increasing the power of the Mann–Whitney test in randomized experiments through flexible covariate adjustment. *Statistics in Medicine*, 34, 1012–1030.

[17] Tsiatis AA, Davidian M, Zhang M, Lu X (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27, 4658–4677.

[18] Zhang M, Tsiatis AA, Davidian M (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64, 707–715.

[19] Thas O, De Neve J, Clement L, Ottoy JP (2012). Probabilistic index models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 74, 623–671.

[20] Chen SX, Qin J, Tang CY (2013). Mann-Whitney test with adjustments to pretreatment variables for missing values and observational study. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 75, 81–102.

[21] Rosenbaum PR, Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

[22] Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, MD.

[23] van der Vaart AW (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

[24] Box GEP, Cox DR (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211-252.

[25] The AFFIRM Investigators (2002). A comparison of rate control and rhythm control in patients with atrial fibrillation. *New England Journal of Medicine*, 347, 1825–1833.

[26] Zhang J, Troendle J, Reddy UM, Laughon SK, Branch DW, Burkman R, Landy HJ, Hibbard JU, Haberman S, Ramirez MM, Bailit JL, Hoffman MK, Gregory KD, Gonzalez-

Quintero VH, Kominiarek M, Learman LA, Hatjis CG, van Veldhuisen P; for the Consortium on Safe Labor (2010). Contemporary cesarean delivery practice in the United States. *American Journal of Obstetrics and Gynecology*, 203, 326.e1–10.

Table 1: Simulation for randomized clinical trials: empirical bias, standard deviation (SD) and relative efficiency (RE) of the empirical and augmented estimators with (approximately) correct and incorrect OR and GPI models (see Section 3 for details).

θ	$\widehat{\theta}$	OR/GPI Model	Bias	SD	RE
		discrete outcome			
0.000	$\widehat{\theta}_{\text{emp}}$		-0.001	0.050	1.00
	$\widehat{\theta}^{\text{or}}_{\text{aug}}$	correct	-0.001	0.042	1.39
	$\widehat{\theta}^{\text{or}}_{\text{aug}}$	incorrect	-0.001	0.044	1.27
	$\widehat{\theta}^{\text{gpi}}_{\text{aug}}$	correct	-0.001	0.042	1.38
	$\widehat{\theta}^{\text{gpi}}_{\text{aug}}$	incorrect	-0.001	0.044	1.26
0.267	$\widehat{\theta}_{\text{emp}}$		0.003	0.046	1.00
	$\widehat{\theta}^{\text{or}}_{\text{aug}}$	correct	0.002	0.038	1.47
	$\widehat{\theta}^{\text{or}}_{\text{aug}}$	incorrect	0.002	0.040	1.33
	$\widehat{\theta}^{\text{gpi}}_{\text{aug}}$	correct	0.002	0.039	1.44
	$\widehat{\theta}^{\text{gpi}}_{\text{aug}}$	incorrect	0.002	0.040	1.32
		continuous outcome			
0.000	$\widehat{\theta}_{\text{emp}}$		0.002	0.053	1.00
	$\widehat{\theta}^{\text{or}}_{\text{aug}}$	correct	0.002	0.039	1.86
	$\widehat{\theta}^{\text{or}}_{\text{aug}}$	incorrect	0.001	0.043	1.52
	$\widehat{\theta}^{\text{gpi}}_{\text{aug}}$	correct	0.002	0.039	1.86
	$\widehat{\theta}^{\text{gpi}}_{\text{aug}}$	incorrect	0.001	0.043	1.52
0.330	$\widehat{\theta}_{\text{emp}}$		-0.003	0.049	1.00
	$\widehat{\theta}^{\text{or}}_{\text{aug}}$	correct	0.000	0.038	1.69
	$\widehat{\theta}^{\text{or}}_{\text{aug}}$	incorrect	0.000	0.042	1.39
	$\widehat{\theta}^{\text{gpi}}_{\text{aug}}$	correct	0.000	0.038	1.70
	$\widehat{\theta}^{\text{gpi}}_{\text{aug}}$	incorrect	0.000	0.042	1.39

Table 2: Simulation for observational studies: empirical bias, standard deviation (SD) and root mean squared error (RMSE) of the regression, IPW and DR estimators with (approximately) correct and incorrect working models (see Section 3 for details).

θ	$\widehat{\theta}$	OR/GPI Model	PS Model	Bias	SD	RMSE
		discrete outcome				
0.000	$\widehat{\theta}_{\text{emp}}$			−0.236	0.046	0.241
	$\widehat{\theta}^{\text{or}}_{\text{reg}}$	correct		−0.001	0.043	0.043
	$\widehat{\theta}^{\text{or}}_{\text{reg}}$	incorrect		−0.059	0.047	0.075
	$\widehat{\theta}^{\text{gpi}}_{\text{reg}}$	correct		−0.010	0.044	0.045
	$\widehat{\theta}^{\text{gpi}}_{\text{reg}}$	incorrect		−0.064	0.047	0.080
	$\widehat{\theta}_{\text{ipw}}$		correct	−0.013	0.092	0.093
	$\widehat{\theta}_{\text{ipw}}$		incorrect	−0.053	0.086	0.101
	$\widehat{\theta}_{\text{dr1}}$	correct	correct	0.000	0.086	0.086
	$\widehat{\theta}_{\text{dr1}}$	incorrect	correct	−0.001	0.110	0.110
	$\widehat{\theta}_{\text{dr1}}$	correct	incorrect	−0.003	0.098	0.098
	$\widehat{\theta}_{\text{dr1}}$	incorrect	incorrect	−0.062	0.100	0.118
	$\widehat{\theta}_{\text{dr2}}$	correct	correct	0.000	0.086	0.086
	$\widehat{\theta}_{\text{dr2}}$	incorrect	correct	−0.002	0.113	0.113
	$\widehat{\theta}_{\text{dr2}}$	correct	incorrect	0.000	0.097	0.097
	$\widehat{\theta}_{\text{dr2}}$	incorrect	incorrect	−0.062	0.100	0.117
0.267	$\widehat{\theta}_{\text{emp}}$			−0.207	0.049	0.213
	$\widehat{\theta}^{\text{or}}_{\text{reg}}$	correct		0.001	0.041	0.041
	$\widehat{\theta}^{\text{or}}_{\text{reg}}$	incorrect		−0.047	0.044	0.064
	$\widehat{\theta}^{\text{gpi}}_{\text{reg}}$	correct		0.006	0.042	0.042
	$\widehat{\theta}^{\text{gpi}}_{\text{reg}}$	incorrect		−0.042	0.045	0.062
	$\widehat{\theta}_{\text{ipw}}$		correct	−0.008	0.068	0.069
	$\widehat{\theta}_{\text{ipw}}$		incorrect	−0.041	0.070	0.081
	$\widehat{\theta}_{\text{dr1}}$	correct	correct	0.000	0.060	0.060
	$\widehat{\theta}_{\text{dr1}}$	incorrect	correct	0.000	0.073	0.073
	$\widehat{\theta}_{\text{dr1}}$	correct	incorrect	0.000	0.068	0.068
	$\widehat{\theta}_{\text{dr1}}$	incorrect	incorrect	−0.044	0.071	0.084
	$\widehat{\theta}_{\text{dr2}}$	correct	correct	0.003	0.066	0.066
	$\widehat{\theta}_{\text{dr2}}$	incorrect	correct	0.000	0.073	0.073
	$\widehat{\theta}_{\text{dr2}}$	correct	incorrect	0.006	0.071	0.071
	$\widehat{\theta}_{\text{dr2}}$	incorrect	incorrect	−0.045	0.073	0.086
		continuous outcome				
0.000	$\widehat{\theta}_{\text{emp}}$			−0.296	0.050	0.300
	$\widehat{\theta}^{\text{or}}_{\text{reg}}$	correct		0.000	0.042	0.042
	$\widehat{\theta}^{\text{or}}_{\text{reg}}$	incorrect		−0.075	0.048	0.089
	$\widehat{\theta}^{\text{gpi}}_{\text{reg}}$	correct		0.001	0.044	0.044
	$\widehat{\theta}^{\text{gpi}}_{\text{reg}}$	incorrect		−0.075	0.049	0.090
	$\widehat{\theta}_{\text{ipw}}$		correct	−0.017	0.101	0.102
	$\widehat{\theta}_{\text{ipw}}$		incorrect	−0.068	0.095	0.117
	$\widehat{\theta}_{\text{dr1}}$	correct	correct	0.005	0.071	0.071
	$\widehat{\theta}_{\text{dr1}}$	incorrect	correct	−0.001	0.102	0.102
	$\widehat{\theta}_{\text{dr1}}$	correct	incorrect	0.005	0.100	0.100
	$\widehat{\theta}_{\text{dr1}}$	incorrect	incorrect	−0.073	0.110	0.132
	$\widehat{\theta}_{\text{dr2}}$	correct	correct	0.005	0.074	0.074
	$\widehat{\theta}_{\text{dr2}}$	incorrect	correct	−0.003	0.106	0.106
	$\widehat{\theta}_{\text{dr2}}$	correct	incorrect	0.005	0.099	0.099
	$\widehat{\theta}_{\text{dr2}}$	incorrect	incorrect	−0.072	0.108	0.130
0.330	$\widehat{\theta}_{\text{emp}}$			−0.282	0.054	0.288
	$\widehat{\theta}^{\text{or}}_{\text{reg}}$	correct		−0.001	0.038	0.038
	$\widehat{\theta}^{\text{or}}_{\text{reg}}$	incorrect		−0.066	0.044	0.080
	$\widehat{\theta}^{\text{gpi}}_{\text{reg}}$	correct		−0.001	0.039	0.039
	$\widehat{\theta}^{\text{gpi}}_{\text{reg}}$	incorrect		−0.066	0.046	0.080
	$\widehat{\theta}_{\text{ipw}}$		correct	−0.012	0.081	0.082
	$\widehat{\theta}_{\text{ipw}}$		incorrect	−0.056	0.078	0.096
	$\widehat{\theta}_{\text{dr1}}$	correct	correct	0.000	0.057	0.057
	$\widehat{\theta}_{\text{dr1}}$	incorrect	correct	−0.001	0.083	0.083
	$\widehat{\theta}_{\text{dr1}}$	correct	incorrect	−0.001	0.071	0.071
	$\widehat{\theta}_{\text{dr1}}$	incorrect	incorrect	−0.065	0.078	0.102
	$\widehat{\theta}_{\text{dr2}}$	correct	correct	0.000	0.056	0.056
	$\widehat{\theta}_{\text{dr2}}$	incorrect	correct	−0.002	0.083	0.083
	$\widehat{\theta}_{\text{dr2}}$	correct	incorrect	−0.002	0.073	0.073
	$\widehat{\theta}_{\text{dr2}}$	incorrect	incorrect	−0.065	0.078	0.102

Table 3: Summary of AFFIRM outcome data: number (percentage) of subjects at each level of Y, by treatment and overall.

T	Y			Row
	0	1	2	Total
0	146 (7.2%)	121 (6.0%)	1760 (86.8%)	2027 (100.0%)
1	175 (8.6%)	126 (6.2%)	1732 (85.2%)	2033 (100.0%)
combined	321 (7.9%)	247 (6.1%)	3492 (86.0%)	4060 (100.0%)

Table 4: Analysis of CSL data: point estimates (PE) and standard errors (SE) of θ for the effect of epidural anagesia on second-stage labor duration, obtained using the regression, IPW and DR methods.

Method	PE	SE
reg-or	0.144	0.019
reg-gpi	0.140	0.019
ipw	0.138	0.019
dr1	0.140	0.019
dr2	0.140	0.019