Published in final edited form as:

Biometrics. 2013 June ; 69(2): 318–327. doi:10.1111/biom.12005.

A Causal Model for Joint Evaluation of Placebo and Treatment-Specific Effects in Clinical Trials

Zhiwei Zhang^{1,*}, Richard M. Kotz¹, Chenguang Wang², Shiling Ruan¹, and Martin Ho¹

¹Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland 20993, U.S.A.

²Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, U.S.A.

Summary

Evaluation of medical treatments is frequently complicated by the presence of substantial placebo effects, especially on relatively subjective endpoints, and the standard solution to this problem is a randomized, double-blinded, placebo-controlled clinical trial. However, effective blinding does not guarantee that all patients have the same belief or mentality about which treatment they have received (or treatmentality, for brevity), making it difficult to interpret the usual intent-to-treat effect as a causal effect. We discuss the causal relationships among treatment, treatmentality and the clinical outcome of interest, and propose a causal model for joint evaluation of placebo and treatment-specific effects. The model highlights the importance of measuring and incorporating patient treatmentality and suggests that each treatment group should be considered a separate observational study with a patient's treatmentality playing the role of an uncontrolled exposure. This perspective allows us to adapt existing methods for dealing with confounding to joint estimation of placebo and treatment-specific effects using measured treatmentality data, commonly known as blinding assessment data. We first apply this approach to the most common type of blinding assessment data, which is categorical, and illustrate the methods using an example from asthma. We then propose that blinding assessment data can be collected as a continuous variable, specifically when a patient's treatmentality is measured as a subjective probability, and describe analytic methods for that case.

Keywords

Blinding; Causal inference; Confounding; Counterfactual; Placebo effect; Potential outcome

1. Introduction

In many therapeutic areas (e.g., neurology, pain, and respiratory disease), treatment evaluation is frequently complicated by a potential placebo effect, that is, the

psychobiological effect of a patient's knowledge or belief of being treated. Here we give the term “placebo” a broad interpretation which includes an inert pill, an inactive device and a sham procedure. It has long been recognized that placebo effects can be substantial, especially on relatively subjective endpoints such as questionnaire-based scores of depression, pain or health-related quality of life. It is, therefore, essential to separate placebo and treatment-specific effects in evaluating a new medical treatment.

The standard solution to this problem is a randomized, double-blinded, placebo-controlled study. It is commonly believed that a simple comparison of treatment groups under such a design represents the causal effect of the investigational treatment, separate from any placebo effect, as long as the blinding procedure is fully effective. However, effective blinding does not guarantee that all patients have the same belief or mentality about which treatment they have received (henceforth abbreviated as treatmentality). A patient who is initially blinded to his or her treatment assignment will nonetheless develop an opinion on the treatment received, based on his or her posttreatment experience (e.g., adverse events, changes in symptoms) as well as certain personal characteristics (e.g., optimism). Such an opinion may vary over time, across patients and possibly between treatment groups, making it difficult to assign a causal interpretation to a simple comparison of outcomes between treatment groups. It is actually possible to measure patient treatmentality by asking a patient to speculate on the treatment received. This has been done in a number of blinded trials for the purpose of evaluating the effectiveness of blinding procedures (e.g., Fuller et al., 1984; Turner et al., 2002; Brinjikji et al., 2010), and the resulting data are usually referred to as blinding assessment data. Several measures of blinding effectiveness have been proposed, including a modified kappa statistic that measures disagreement (James et al., 1996) and a pair of indices (one for each treatment group) based on the proportion of correct answers beyond chance (Bang, Ni, and Davis, 2004). Bang et al. (2010) review and compare the available methods and propose a blinding assessment protocol. So far, there has been no discussion on how to incorporate blinding assessment data into a treatment comparison for the clinical outcome of interest.

Successful integration of blinding assessment and treatment evaluation requires an appropriate causal framework. While it may be tempting to adjust for treatmentality as a confounder, this is not appropriate because treatmentality is a postrandomization variable. In this article, we propose a causal model that extends the classical treatment effect model of Rubin (1974) by incorporating a placebo effect, thereby allowing joint evaluation of placebo and treatment-specific effects. Under the proposed model, we show that, without relevant information about patient treatmentality, the usual treatment comparison does not have a clear causal interpretation even with perfect blinding, which we take to mean that treatmentality is independent of treatment assignment. We note that each treatment group can be regarded as a separate observational study with a patient's treatmentality playing the role of an uncontrolled exposure. This observation allows us to borrow techniques from the causal inference literature and develop methods for estimating placebo and treatment-specific effects using blinding assessment data. We apply this approach to a standard categorical format of blinding assessment, where patients are asked to indicate which treatment they believe they have received, with the option of answering “don't know” if they do not have a strong belief. We also suggest that treatmentality can be measured

continuously as a subjective probability (of having received the experimental treatment), and describe analytic methods for that case.

The proposed approach differs profoundly from the latent variable modeling approach of Eickhoff (2008) for joint estimation of treatment and placebo effects, where the placebo effect is treated as a source of measurement error and separated from the treatment effect using modeling assumptions. The latter approach does not involve blinding assessment, and is not designed to answer our research questions concerning the joint effects of treatment and treatmentality.

The rest of this article is organized as follows. In the next section, we present a causal model for the joint effects of treatment and treatmentality, discuss its implications on treatment comparison, and draw connections with the causal inference literature. In Section 3, we describe the standard format of blinding assessment data and develop methods for estimating placebo and treatment-specific effects. In Section 4, the methods are illustrated with an asthma-related quality-of-life example and evaluated in a simulation study based on the same example. In Section 5, we consider continuous treatmentality and develop the associated methods. The article concludes with a discussion in Section 6.

2. Rationale

We start by formulating the research questions in terms of potential outcomes. Suppose an experimental treatment is to be compared with placebo with respect to a clinical outcome that is potentially subject to a placebo effect. For a generic patient in the target population, let $Y(t, s)$ denote the potential outcome that will realize if the patient receives treatment t (0 for placebo; 1 for experimental) and believes at the time of evaluation that he or she has received treatment s . Note that we assume a fixed time point for outcome evaluation and do not consider the longitudinal setting. We recognize that a patient's treatmentality can be quite complex, and do not assume that it is dichotomous. We will, however, focus on two possible values ($s = 0, 1$) in defining the causal effects of interest.

For any fixed value s , the difference $Y(1, s) - Y(0, s)$ represents the effect of the experimental treatment on an individual patient. Let $\mu_t s = E\{Y(t, s)\}$, $t, s = 0, 1$. Then the population-average effect of the treatment can be defined as

$$\delta_{.s} = \mu_{1s} - \mu_{0s} \quad (1)$$

for any s , or perhaps as a weighted average over s . Switching the roles of t and s leads to analogous definitions of placebo effects, namely,

$$\delta_{t.} = \mu_{t1} - \mu_{t0} \quad (2)$$

In addition, one may be interested in a possible interaction between t and s , defined as

$$\tau = \delta_{.1} - \delta_{.0} = \delta_{1.} - \delta_{0.}.$$

While some of these quantities may be of greater interest than others in a particular application, we will develop general methodology for estimating the μ_{ts} , from which estimates of all effects and the interaction can be derived easily.

The causal effects defined by (1) and (2) have clear interpretations in clinical practice. In practice, treatment assignment is usually known to the patient, and the resulting outcome is either $Y(1, 1)$ (if treated) or $Y(0, 0)$ (if untreated). Thus, it makes sense to consider the difference $Y(1, 1) - Y(0, 0)$ as the total effect of applying the experimental treatment to an individual patient. The average total effect, $\delta_{..} = \mu_{11} - \mu_{00}$, can be written as the sum of a treatment effect and a placebo effect:

$$\delta_{..} = \delta_{0.} + \delta_{.1} = \delta_{0.} + \delta_{1.}$$

Each decomposition results from raising t and s from 0 to 1 sequentially (s first in the middle expression; t first in rightmost expression). If there is no interaction between t and s , then the two sequences lead to identical effects.

If treatmentality is considered a mediator, the effects defined by (1) and (2) can be seen as controlled direct and indirect effects (e.g., Robins and Greenland, 1992; Rubin, 2004; VanderWeele, 2008). Related to these concepts are “natural” direct and indirect effects as well as definitions based on principal stratification. We do not consider these related concepts, which involve potential outcomes for treatmentality (indexed by treatment) following well-defined distributions. Although such potential outcomes may exist in a clinical study, they are more characteristic of the study than of the target population. In contrast, the controlled effects defined by (1) and (2) are truly population quantities in that they are based on clinical practice in a given population as opposed to the characteristics of a clinical study.

Suppose now that a randomized clinical trial is conducted to answer our research questions. Let T denote the treatment assigned to a study subject; thus T is a Bernoulli variable independent of all baseline variables. Without considering noncompliance, we assume that T is also the actual treatment given to the subject. We assume that the study is designed as blinded, that is, patients are not informed of their treatments. To crystalize the main ideas, we start here with the simplistic assumption that every patient has a strong belief regarding his or her treatment and is willing to express it. (More realistic ways to characterize and measure patient treatmentality will be considered after this section.) We denote this belief by S , a Bernoulli variable, and write $Y = Y(T, S)$ for the actual outcome. Implicit in the latter notation is the assumption of consistency or stable unit treatment value.

In this setting, treatment evaluation is typically based on an intent-to-treat (ITT) analysis, which estimates the ITT effect $E(Y|T=1) - E(Y|T=0)$. However, we show in Web Appendix A that, without any information about S , the ITT effect is hardly interpretable as a causal effect of the new treatment under the causal model described earlier. Moreover, an ITT analysis does not help with understanding the placebo effect. To answer our research questions, there is clearly a need for some adjustment, perhaps using information about S .

If we could assume that S is conditionally independent of the potential outcomes $Y(t, s)$ given T , written

$$S \perp Y(t, s) | T=t, \quad t, s=0, 1, \quad (3)$$

then we have $E(Y|T=t, S=s) = \mu_{ts}$, which implies that μ_{ts} can be estimated by averaging Y among subjects with $T=t, S=s$. While convenient to work with, assumption (3) is often unrealistic because S does depend on a patient's personal characteristics (e.g., optimism) and posttreatment experience (e.g., adverse events and changes in symptoms). If some of these determinants of S are also related to the outcome of interest, then assumption (3) is highly questionable. The problem is also known as confounding in the literature on causal inference in observational studies. Indeed, each treatment group ($T=t$) can be considered a separate observational study with S as an uncontrolled exposure and $\{Y(t, s) : s=0, 1\}$ as the potential outcomes. This connection allows us to estimate the μ_{ts} by adapting existing techniques for causal inference.

It is important to have available a set of confounders, denoted by $\mathbf{X}$, that are sufficient to explain any association between S and the potential outcomes. Formally, we assume that

$$S \perp Y(t, s) | \mathbf{X}, T=t, \quad t, s=0, 1, \quad (4)$$

which is often referred to as no unmeasured confounding or strong ignorability (Rosenbaum and Rubin, 1983a). This assumption is more realistic than assumption (3) in that S may now depend on potential outcomes through a vector of covariates. In practice, $\mathbf{X}$ may be chosen as a set of baseline characteristics and/or posttreatment measurements that are associated with both treatment and the outcome of interest. Recall that the confounding of concern here is with respect to S , not T , in a treatment group defined by $T=t$; thus, even though $\mathbf{X}$ may include posttreatment variables, they can be considered confounders for S in the sense that they precede and predict S . We do require that $\mathbf{X}$ be fully objective and not itself subject to a placebo effect. Also key to our approach is the positivity assumption that, with probability 1,

$$P(S=s | \mathbf{X}, T=t) > 0, \quad t, s=0, 1. \quad (5)$$

The essence of assumption (5) is that both states of treatment ($s=0, 1$) are possible for all subjects in both treatment groups with different characteristics and experiences. Unlike assumption (4), which is not testable with the observed data, assumption (5) can and should be checked with the data.

Together, assumptions (4) and (5) allow the μ_{ts} to be identified nonparametrically and estimated using standard methods (e.g., van der Laan and Robins, 2003; Bang and Robins, 2005), provided S is binary and measured accurately. The latter assumption is unrealistic, however. In the next section, we develop practical methods based on realistic assumptions about the nature of S and how it is measured.

3. Methodology

In a standard procedure for blinding assessment, patients are asked to indicate which treatment they believe they have received, with the option of answering “don't know” if they do not have a strong belief. This 3-level categorization is more realistic than the dichotomization assumed in Section 2, and is commonly used in practice. The goal of this section is to develop methods for estimating the μ_{ts} using blinding assessment data in this 3-level format. Also in current use is a 5-level format where a patient has to indicate the strength of his or her belief as “strong” or “moderate” when a belief is expressed (see Bang et al., 2010, for illustrations and a comparison of the 5-level and 3-level formats). Our methods for the 3-level format extend easily to the 5-level format, which therefore will not be discussed in this article.

Under the 3-level design, the sample space for S may be defined as $\{0, u, 1\}$, where u represents “unknown.” An analytical challenge in blinding assessment is that S may not be measured accurately. With the 3-level design, a major concern is that patients may be tempted to conform to a perceived expectation of remaining blinded and answer “don't know” even when they have a strong belief. To account for this possible misclassification, let S^* denote the answer given by a patient and let R indicate the truthfulness of the answer (1 if true; 0 if false). Assuming that a false u is the only possible false answer, we then have

$$S^* = RS + (1 - R)u \quad (6)$$

with u treated as an unknown number in arithmetic operations. Thus, for $s = 0, 1$, we have that $S^* = s$ if and only if $S = s$ and $R = 1$. The observed data can be written as $(T_i, \mathbf{X}_i, S_i^*, Y_i), i=1, \dots, n$, which we conceptualize as independent copies of $(T, \mathbf{X}, S^*, Y)$.

We now consider estimation of μ_{ts} for fixed $t, s = 0, 1$. We do not consider $Y(t, s)$ or μ_{ts} for $s = u$, which are irrelevant to the causal effects defined in Section 2. In addition to assumptions (4) and (5), we assume that

$$R \perp Y(t, s) \mid \mathbf{X}, T=t, S=s \quad (7)$$

and that, with probability 1,

$$P(R=1 \mid \mathbf{X}, T=t, S=s) > 0 \quad (8)$$

in order to deal with possible misclassification into the “don't know” category. Under the misclassification mechanism given by (6), the potential outcome $Y(t, s)$ is knowingly observed if and only if $T = t$ and $S^* = s$. Note that $Y(t, s)$ can be observed unknowingly if $T = t, S = s$, and $S^* = u$ (because $R = 0$). Now, assumptions (4) and (7) together imply that

$$P\{S^* = s \mid \mathbf{X}, T=t, Y(t, s)\} = P(S^* = s \mid \mathbf{X}, T=t),$$

which is essentially a missing-at-random condition in the sense of Rubin (1976). Assumptions (5) and (8) together further ensure that

$$P(S^* = s | \mathbf{X}, T=t) = P(S=s, R=1 | \mathbf{X}, T=t) > 0$$

with probability 1, and hence that μ_{ts} is identified nonparametrically.

It is straightforward to estimate μ_{ts} using an outcome regression (OR) approach based on the regression function $m(T, S^*, \mathbf{X}) = E(Y|T, S^*, \mathbf{X})$, where S^* is treated as a categorical variable. For $s = 0, 1$ (but not u), we have

$$m(t, s, \mathbf{X}) = E\{Y(t, s) | T=t, \mathbf{X} = \mathbf{X}\},$$

which follows from the definition of S^* together with assumptions (4) and (7). Now μ_{ts} can be identified as

$$E\{Y(t, s)\} = E\{Y(t, s) | T=t\} = E[E\{Y(t, s) | T=t, \mathbf{X}\} | T=t] = E\{m(t, s, \mathbf{X}) | T=t\}, \quad (9)$$

and estimated by an empirical analogue of the last expression in (9). Let the regression function be parameterized as $m(t, s, \mathbf{x}; \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a finite-dimensional parameter, and denote by $\hat{\boldsymbol{\beta}}$ an estimate of $\boldsymbol{\beta}$, which may be obtained by (iteratively reweighted) least squares, depending on the model. Then the OR estimator of μ_{ts} is

$N_t^{-1} \sum_{i=1}^n I(T_i=t) m(t, s, \mathbf{X}_i; \hat{\boldsymbol{\beta}})$, where $I(\cdot)$ is the indicator function and $N_t = \sum_{i=1}^n I(T_i=t)$ is the size of the t th treatment group. The far right side of (9) reduces to the marginal expectation $E\{m(t, s, \mathbf{X}; \boldsymbol{\beta})\}$ if $\mathbf{X}$ consists of baseline characteristics only, in which case a more efficient estimator of μ_{ts} is given by $n^{-1} \sum_{i=1}^n m(t, s, \mathbf{X}_i; \hat{\boldsymbol{\beta}})$.

Next, we derive an inverse probability weighting (IPW) method based on the fact that

$$\mu_{ts} = E\left\{ \frac{I(S^* = s)Y}{P(S^* = s | T=t, \mathbf{X})} | T=t \right\},$$

which follows from a conditioning argument (Horvitz and Thompson, 1952; Robins, Rotnitzky and Zhao, 1994). This method can be implemented by specifying a propensity score (PS) model $\pi_s(T, \mathbf{X}; \gamma) = P(S^* = s | T, \mathbf{X})$, where π is a known function and γ an unknown finite-dimensional parameter. Denote by $\hat{\gamma}$ the maximum likelihood estimate of γ .

Then the IPW estimator of μ_{ts} is given by $N_t^{-1} \sum_{i=1}^n I(T_i=t, S_i^*=s) Y_i / \pi_s(t, \mathbf{X}_i; \hat{\gamma})$.

An augmented IPW (AIPW) estimator of μ_{ts} can be obtained as

$$\frac{1}{N_t} \sum_{i=1}^n I(T_i=t) \left\{ \frac{I(S_i^*=s)Y_i}{\pi_s(t, \mathbf{X}_i; \hat{\gamma})} - \frac{I(S_i^*=s) - \pi_s(t, \mathbf{X}_i; \hat{\gamma})}{\pi_s(t, \mathbf{X}_i; \hat{\gamma})} m(t, s, \mathbf{X}_i; \hat{\beta}) \right\},$$

which is well known to be doubly robust with respect to the OR and PS models and locally efficient when both models are correct (e.g., van der Laan and Robins, 2003; Bang and Robins, 2005). Additional doubly robust estimators of μ_{ts} could be derived by drawing upon recent developments in doubly robust estimation (see van der Laan and Rubin, 2006; Tan, 2010, and the references therein).

The estimators described so far do not involve any constraints among the μ_{ts} . If one is confident that there is no interaction between t and s , in the sense that $0 = \tau = \mu_{11} - \mu_{10} - \mu_{01} + \mu_{00}$, an empirical likelihood method can then be used to enforce the constraint and enhance the efficiency of the estimators (Qin and Lawless, 1994). The empirical likelihood approach involves multidimensional constrained maximization in our setting, and thus may be difficult to implement in practice. Here, we propose a pseudo-empirical likelihood (PEL) approach that is similar in spirit but easier to implement. Each of the unconstrained

estimators proposed earlier can be written as $\hat{\mu}_{ts} = n^{-1} \sum_{i=1}^n h_n(T_i, \mathbf{X}_i, S_i^*, Y_i)$ for some function h_n which may depend on $\hat{\beta}$, $\hat{\gamma}$, and N_t . The corresponding PEL estimator is given by

$\tilde{\mu}_{ts}(\mathbf{w}) = \sum_{i=1}^n w_i h_n(T_i, \mathbf{X}_i, S_i^*, Y_i)$, where the weights $\mathbf{w} = (w_1, \dots, w_n)$ are chosen to

maximize the empirical log-likelihood $\sum_{i=1}^n \log w_i$ subject to the following constraints: $w_i \geq 0 \forall i$, $\sum_{i=1}^n w_i = 1$, and $0 = \mu_{11}(\mathbf{w}) - \mu_{10}(\mathbf{w}) - \mu_{01}(\mathbf{w}) + \mu_{00}(\mathbf{w})$. The maximizer can be found easily using a bisection method (Owen, 2001).

All of the aforementioned estimators are asymptotically normal under regularity conditions, with closed-form variance estimators given in Web Appendix B. Alternatively, bootstrap standard errors and confidence intervals can be used for inference. Some modeling issues are discussed in Web Appendix C. R code for the proposed methods is provided in Web Appendix D.

4. Numerical Results

We now illustrate the methods with a randomized, double-blinded, multicenter clinical trial comparing an investigational device with a sham control for treating severe persistent asthma in adults. About 300 patients were enrolled in multiple countries and randomized at a 2 : 1 ratio to receive an active treatment with the new device or a sham treatment with the same device (power off). Both treatments were delivered in a bronchoscopy procedure under local anesthesia. Both the patients and the assessors (who evaluated the patients following the treatment) were blinded, although the treating physicians could not be blinded. Follow-up evaluations were performed at multiple time points (out to 1-year posttreatment) to collect information on adverse events, medications, treatmentality, asthma-related symptoms and quality of life, as well as certain clinical measurements such as morning peak respiratory flow (MPRF). Treatmentality was measured in the 3-level format described in Section 3. Asthma-related symptoms were summarized into a score that ranges between 0 and 12, with

higher scores indicating better (less severe) symptoms. Asthma-related quality of life was measured by the Asthma Quality of Life Questionnaire (AQLQ), which leads to a score that ranges from 1 to 7, with higher values indicating better quality of life.

The outcome measure of primary interest in our analysis is the change in the AQLQ score from baseline to 6 months posttreatment, which is clinically important but relatively subjective, raising concerns about a possible placebo effect. Our analysis is restricted to a subset of 128 patients in the United States due to apparent heterogeneity between U.S. and non-U.S. sites and within the latter group of sites. As possible confounders for treatmentality, we consider gender, age, number of years with asthma (NYA), baseline AQLQ score, and baseline and 6-month symptom scores and MPRF measurements. Our OR model is a global model (in the sense of Web Appendix C) relating Y linearly to T , S^* , baseline AQLQ score, baseline and 6-month symptom scores, and an interaction between T and baseline AQLQ score, with or without an interaction between T and S^* (to be discussed later). Our PS model is a two-stage model described in Web Appendix C. The first stage, for $P(S^* = u|T, X)$, is a logistic regression model with T and NYA as the only linear terms. The second stage, for $P(S^* = 1|T, X, S^* = u)$, is another logistic model which includes T , NYA, baseline AQLQ score, and baseline and 6-month symptom scores, with an interaction between T and baseline AQLQ score. Both the OR and the PS models were constructed systematically, starting with univariate analyses before considering possible interactions.

The data are analyzed using the OR, IPW, and AIPW methods described in Section 3, a partially adjusted (PA) method which estimates μ_{ts} as $E(Y|T = t, S = s)$, and an unadjusted method which estimates μ_{ts} as $E(Y|T = t)$. The methods are used to estimate the μ_{ts} with and without allowing a nonzero interaction between t and s (i.e., $\tau \neq 0$). In the present context, it is easy to see that $\tau = 0$ if and only if T interacts with S^* in the OR model. Therefore, an interaction between T and S^* is included in the OR model when and only when a nonzero τ is allowed. When τ is assumed to be zero, the PEL approach of Section 3 is used to enforce the constraint for all methods but the unadjusted one, which ignores any placebo effect and always estimates τ as zero, regardless of what is assumed about τ .

Table 1 presents point estimates and standard errors (based on 1000 bootstrap samples) of the μ_{ts} and the derived quantities (causal effects and interaction) for the example. Consider first unconstrained estimation without assuming $\tau = 0$. The unadjusted analysis yields an ITT effect estimate of 0.61, which is significantly different from 0, but is uninformative about any placebo effect and a possible interaction. The PA analysis, which utilizes treatmentality data but fails to adjust for confounding, gives similar estimates for the treatment effect and suggests the presence of a substantial placebo effect. The three methods (OR, IPW, and AIPW) that adjust for confounding produce fairly consistent results, which together suggest that the treatment effect is stronger at $s = 0$ and that the placebo effect is stronger at $t = 0$. Although the interaction test is not significant for any method, which is not surprising given the moderate sample size, there appears to be a high level of agreement among the last three methods in suggesting the presence of a sizable negative interaction. One possible explanation for the negative interaction (if true) might be that treatment and treatmentality share some causal pathways. If some common pathways are saturated by one factor (treatment or treatmentality), then the other factor will not have any additional effect

through the same pathways, hence the negative interaction. For illustrative purposes, we also analyze the data under the assumption that $\tau = 0$, even though this is not clearly the case. This assumption has no effect on the unadjusted method, as noted earlier, but leads to modified estimates that satisfy $\delta_0 = \delta_1$ and $\delta_0 = \delta_1$ for the other methods. The constrained estimates of the treatment effect are strikingly similar to each other, with the possible exception of the AIPW estimate. The constrained estimates of the placebo effect are less similar. Together, the results in Table 1 suggest the presence of a positive treatment effect, a positive placebo effect and a negative interaction between the two, with decreasing levels of confidence.

The methods are evaluated and compared in a simulation study based on the same example. Specifically, we focus on the same set of 128 U.S. patients (except in one case, to be described later) with the original values of (T, X) , and generate values of (S, Y) randomly under the PS and OR models described earlier with parameter values estimated from the previous analyses. As before, the OR model may or may not include an interaction between T and S^* , depending on whether τ is allowed to differ from zero. In each scenario, 10^4 datasets are simulated and analyzed using the same methods as in Table 1, with working PS and OR models that may or may not be the same as the models for data generation. Table 2 compares the methods in terms of bias and standard deviation under correct model specification. Clearly, the unadjusted and PA methods produce biased estimates, though the bias is smaller for some estimands. In contrast, the OR and AIPW methods are virtually unbiased for all estimands. The IPW method shows some bias at $n = 128$, especially under the constraint $\tau = 0$. To see if this is a small-sample bias, we increase the sample size to 640 (five times the original size) by adding four replicates of the original patients. As shown in the lower half of Table 2, the bias of the IPW method becomes negligible at the larger sample size. As predicted by theory, OR is the most efficient among the three proposed methods, and IPW the least efficient, when both working models are correct.

The methods are also compared when either or both of the working PS and OR models are misspecified. For both models, misspecification results from omitting an important covariate (baseline AQLQ). Table 3 shows the results for the unconstrained case (i.e., not assuming $\tau = 0$) only, as the constrained case follows the same pattern (except that a larger sample size is needed for the IPW method to be practically unbiased). In Table 3, the OR method is severely biased when the OR model is misspecified, as is the IPW method when the PS model is misspecified. In each case, the AIPW method remains practically unbiased, as expected. When both models are misspecified, all methods are biased. In terms of variability, the OR, IPW, and AIPW methods compare to each other in much the same way as in Table 2. Thus, it appears that the OR and AIPW methods have unique advantages, and the choice between them is largely a bias-variance trade-off. The AIPW method is more robust in terms of bias, while the OR method tends to be more efficient. No clear advantage has been observed for the IPW method.

5. Continuous Treatmentality

A common concern about the blinding assessment data described in Section 3 is that the 3-level format, or even the 5-level format, cannot capture all relevant information about a

patient's treatmentality. Furthermore, the “don't know” category creates some ambiguity and complicates the analysis and interpretation of data. In this section, we propose a new design for blinding assessment on an (approximately) continuous scale and develop methods for estimating causal effects under the new design.

It seems natural to think of treatmentality as a subjective probability with range $[0, 1]$. The inclusion of 0 and 1, which correspond to absolute certainty, may be somewhat unrealistic. For our purpose, however, the difference between near certainty and absolute certainty is not important. The values $s = 0, 1$ are useful in formulating research questions and convenient to work with in data analysis. The potential outcome $Y(t, s)$ is now defined for any $(t, s) \in \{0, 1\} \times [0, 1]$, and we write $\mu(t, s) = E\{Y(t, s)\}$.

As a subjective probability, treatmentality could be measured directly by asking a study participant about his or her perceived likelihood of having been treated, given his or her experience. Alternatively, the question could be framed in a hypothetical betting scheme, often used to elicit a subjective probability. Either way, the instrument should be designed carefully to ensure that an average patient can understand the question correctly and respond rationally. The instrument used by Brinjikji et al. (2010), with a forced choice qualified by a numerical confidence level, might be a good starting point. Also, there is no guarantee that a forced numerical answer (instead of “don't know”) is necessarily truthful. The truthfulness issue, which may be considered a missing data problem under the 3-level design, could persist under the new design as a measurement error problem. These issues should be considered carefully in designing the instrument, but will not be discussed further in this article, which focuses on statistical analysis. In what follows we assume that S is measured accurately, so that the observed data now consist of the independent copies (T_i, X_i, S_i, Y_i) , $i = 1, \dots, n$.

If we are only interested in estimating the causal effects involving $s = 0, 1$ and no intermediate values, and if S has a positive probability at each extreme value, then we could define S^* as u if $0 < S < 1$ and as S otherwise, and employ the methods described in Section 3 under assumptions (4) and (5). Sometimes, it is important to also understand how $\mu(t, s)$ changes as s goes from 0 to 1. The rest of this section is concerned with estimation of $\mu(t, s)$ as a function of $(t, s) \in \{0, 1\} \times [0, 1]$.

The causal inference perspective of Section 2 applies to the present situation, except that S is now a quantitative exposure. We assume that assumption (4) holds for $(t, s) \in \{0, 1\} \times [0, 1]$ and that

$$f(s|t, \mathbf{X}) > 0 \quad \forall (s, t, \mathbf{X}), \quad (10)$$

where $f(\cdot|t, \mathbf{x})$ is the conditional density of S given $T = t$ and $\mathbf{X} = \mathbf{x}$ with respect to some measure ν on $[0, 1]$. It is necessary to consider a general measure here because S may be continuous on $(0, 1)$ and discrete at 0 and 1. The function $f(s|t, \mathbf{x})$ generalizes the PS for a binary exposure and is known as the propensity function (PF) (Imai and van Dyk, 2004). Like assumption (5), assumption (10) requires that all levels of treatmentality be possible for all subjects in each treatment group. This assumption can and should be verified empirically,

as its violation can have serious consequences (Moore et al, 2012). Assumptions (4) and (10) together ensure non-parametric identification of $\mu(t, s)$ for all (t, s) (Zhang et al., 2012).

Under an OR model given by $m(t, s, \mathbf{x}; \beta) = E(Y|T = t, S = s, X = \mathbf{x})$, it can be argued as before that

$$\mu(t, s) = E \{ m(t, s, \mathbf{X}; \beta) | T = t \}. \quad (11)$$

If $\hat{\beta}$ is a consistent estimate of β , then $\mu(t, s)$ is estimated consistently by

$$N_t^{-1} \sum_{i=1}^n I(T_i = t) m(t, s, \mathbf{X}_i; \hat{\beta}).$$

In the present context, the IPW method involves a marginal structural model (MSM) for $\mu(t, s)$, which may be parametric or nonparametric, and a PF model $f(s|t, \mathbf{x}; \alpha)$, which is usually a parametric model for the conditional density of S given $(T, \mathbf{X})$. For a parametric PF model,

the parameter α can be estimated by maximizing the likelihood $\prod_{i=1}^n f(S_i|T_i, \mathbf{X}_i; \alpha)$, and the resulting estimate will be denoted by $\hat{\alpha}$. Following the IPW approach of Robins, and Brumback (2000), $\mu(t, s)$ can be estimated by fitting the MSM (i.e., regressing Y_i on (T_i, S_i)) with each subject weighted by $W(S_i|T_i)/f(S_i|T_i, \mathbf{X}_i; \hat{\alpha})$, where $W(s|t) > 0$ is chosen to stabilize the weight when $f(S_i|T_i, \mathbf{X}_i; \hat{\alpha})$ in the denominator is very small for some subjects. A common choice for W is an estimated density of S given T .

For a parametric MSM, an AIPW method that is doubly robust and locally efficient is available from van der Laan and Robins (2003). However, the double robustness of the AIPW method is relevant only if the MSM and the OR model are compatible in the sense that equation (11) holds for some β (so that the MSM does not rule out the OR model a priori). When $\mu(t, s)$ is modeled parametrically, equation (11) can hold in special cases (i.e., linear models and some log-linear models) but not in general (Zhang et al., 2012).

Moreover, the AIPW method involves an integral in the estimating function and may be difficult to implement. More readily available is the weighted regression approach of Zhang et al. (2012), which simply fits the OR model with the same IPW weight defined in the last paragraph. Denote by $\hat{\beta}^*$ the resulting estimate of β . Then the weighted regression estimator

of $\mu(t, s)$ is given by $N_t^{-1} \sum_{i=1}^n I(T_i = t) m(t, s, \mathbf{X}_i; \hat{\beta}^*)$. The weighted regression method is doubly robust for estimating a parametric MSM with a compatible OR model (the same situations in which the AIPW method is doubly robust), and also for estimating a nonparametric MSM with a compatible OR model (which could be specified as a generalized additive model with a nonparametric component for S). Further information about the weighted regression method (and other possible methods) can be found in Zhang et al. (2012).

6. Discussion

As the term indicates, blinding assessment data have historically been collected to assess the effectiveness of blinding procedures. This article proposes to use blinding assessment data for a different purpose (i.e., estimation of placebo and treatment-specific effects under a

joint causal model). The proposed approach is complementary to the existing methods for blinding assessment. On the one hand, our proposal is not explicitly concerned with blinding success and does not require perfect blinding. Indeed, one could argue that partial unblinding is inevitable with an effective treatment because patients learn from their experiences. Without further evidence of faulty experimentation, partial unblinding alone need not be considered a fatal flaw of a clinical trial. On the other hand, unblinding has profound implications that are not limited to the placebo effect. As noted by the associate editor, unblinding can lead to behavioral changes that eventually result in drop-out, noncompliance and/or other protocol violations. To protect the overall quality and integrity of a clinical trial, it remains important to proactively strive for blinding success in the design and conduct of the trial, carefully evaluate the extent of blinding success in the analysis stage (using blinding assessment data), and thoroughly identify and examine the reasons for any (partial) unblinding that did occur.

The quality of blinding assessment data is essential to valid and meaningful inference based on the proposed methods. For the standard (3- and 5-level) formats, Bang et al. (2010) provide practical recommendations on how to ensure data quality. Here, we suggest a new design for measuring treatmentality continuously as a subjective probability. The new design has the potential of recovering some information lost in the “don't know” category of the standard formats, although its implementation can be challenging and will require some further thoughts.

The proposed approach requires the availability of a set of confounders for treatmentality satisfying assumption (4) (i.e., no unmeasured confounders). Ideally, this assumption should be addressed by a thoughtful design which ensures that all relevant information be collected. This includes certain personal characteristics such as optimism or wishful thinking, which may be reflected in baseline treatment guesses and predicted outcomes (Turner et al., 2002), as well as relevant post-treatment experiences (e.g., adverse events, changes in symptoms). It is important to recognize that assumption (4) is not testable with the observed data and must be based on substantive knowledge. If there are serious doubts about this assumption in the analysis stage, a sensitivity analysis could be performed to assess the impact of unmeasured confounders (e.g., Rosenbaum and Rubin, 1983b; Lin, Psaty and Kronmal, 1998; Brumback et al., 2004).

Application of the proposed methods does not require the blinding procedure to be perfect or meet a certain performance criterion. It does require the positivity assumption (5) or its continuous version (10) in order to identify μ_{ts} or $\mu(t, s)$, respectively. Thus, for example, if all patients in the control group guess correctly (i.e., $S = 0$ whenever $T = 0$), then μ_{01} is not empirically identifiable.

While developed for placebo-controlled trials, the proposed methodology can certainly be used in active-controlled trials upon redefining the placebo effect as the psychobiological effect of the knowledge or belief of receiving one treatment versus the other (e.g., new treatment versus standard of care). The latter effect, though less common than the usual placebo effect, may arise from perceived advantages (e.g., new technology) of the new treatment.

The proposed methodology could be extended in several directions. Although we have focused on the simple situation of a single time point for outcome evaluation, the general approaches of van der Laan and Robins (2003) and Bang and Robins (2005), after suitable modifications, could be used to analyze longitudinal data measured at multiple time points. Also, with focus on blinding, we have not discussed other complications such as noncompliance, drop-out and protocol violations, which may occur for a variety of reasons including adverse events, changes in symptoms and, to complicate things even further, (partial) unblinding. While these do not appear to be major issues in our example, they can certainly create problems in other trials. It will be of interest to address these complications in the context of a suitable application.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the editor, the associate editor and a reviewer for insightful and constructive comments that have greatly improved the article. The views expressed in this article do not represent the official position of the U.S. Food and Drug Administration.

References

- Bang H, Flaherty SP, Kolahi J, Park J. Blinding assessment in clinical trials: A review of statistical methods and a proposal of blinding assessment protocol. *Clinical Research and Regulatory Affairs*. 2010; 27:42–51.
- Bang H, Ni L, Davis CE. Assessment of blinding in clinical trials. *Controlled Clinical Trials*. 2004; 25:143–156. [PubMed: 15020033]
- Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005; 61:962–972. [PubMed: 16401269]
- Brinjikji W, Comstock BA, Heagerty PJ, Jarvik JG, Kallmes DF. Investigational vertebroplasty efficacy and safety trial: Detailed analysis of blinding efficacy. *Radiology*. 2010; 257:219–225. [PubMed: 20851942]
- Brumback BA, Hernan MA, Haneuse SJ, Robins JM. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*. 2004; 23:749–767. [PubMed: 14981673]
- Eickhoff JC. Placebo effect-adjusted assessment of quality of life in placebo-controlled clinical trials. *Statistics in Medicine*. 2008; 27:1387–1402. [PubMed: 18219702]
- Fuller RK, Williford WO, William O, Lee KK, Derman R. Veterans Administration cooperative study of disulfiram for the treatment of alcoholism: Study design and methodological considerations. *Controlled Clinical Trials*. 1984; 5:263–273. [PubMed: 6386330]
- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47:663–685.
- Imai K, van Dyk DA. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*. 2004; 99:854–866.
- James KE, Bloch DA, Lee KK, Kraemer HC, Fuller RK. An index for assessing blindness in a multi-centre clinical trial: Disulfiram for alcohol cessation—a VA cooperative study. *Statistics in Medicine*. 1996; 15:1421–1434. [PubMed: 8841652]
- Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 1998; 54:948–963. [PubMed: 9750244]
- Moore KL, Neugebauer R, van der Laan MJ, Tager IB. Causal inference in epidemiological studies with strong confounding. *Statistics in Medicine*. 2012; 31:1380–1404. [PubMed: 22362629]

- Owen, AB. Empirical likelihood. Boca Raton, FL: Chapman & Hall; 2001.
- Qin J, Lawless J. Empirical likelihood and general estimating equations. *Annals of Statistics*. 1994; 22:300–325.
- Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992; 3:143–155. [PubMed: 1576220]
- Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000; 11:550–560. [PubMed: 10955408]
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983a; 70:41–55.
- Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B*. 1983b; 45:212–218.
- Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974; 66:688–701.
- Rubin DB. Inference and missing data. *Biometrika*. 1976; 63:581–592.
- Rubin DB. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*. 2004; 31:161–170.
- Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*. 2010; 97:661–682.
- Turner JA, Jensen MP, Warms CA, Cardenas DD. Blinding effectiveness and association of pretreatment expectations with pain improvement in a double-blind randomized controlled trial. *Pain*. 2002; 99:91–99. [PubMed: 12237187]
- van der Laan, MJ.; Robins, JM. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer-Verlag; 2003.
- van der Laan MJ, Rubin DB. Targeted maximum likelihood learning. *The International Journal of Biostatistics*. 2006; 2(1) Article 11. 10.2202/1557-4679.1043
- VanderWeele TJ. Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters*. 2008; 78:2957–2962.
- Zhang Z, Zhou J, Cao W, Zhang J. Causal inference with a quantitative exposure. *Statistical Methods in Medical Research*. 2012 in press.

Table 1

Analysis of quality-of-life data in Section 4: point estimates and standard errors (based on 1000 bootstrap samples) for the μ_{ts} and the derived quantities, obtained using the unadjusted (UA), PA, OR, IPW, and AIPW methods, with and without allowing a nonzero interaction between t and s (i.e., $\tau = 0$)

Parameter	Point estimate					Standard error				
	UA	PA	OR	IPW	AIPW	UA	PA	OR	IPW	AIPW
Not assuming $\tau = 0$										
μ_{00}	0.97	0.39	0.39	0.47	0.28	0.16	0.27	0.29	0.39	0.56
μ_{01}	0.97	1.18	1.04	1.20	1.09	0.16	0.24	0.21	0.34	0.27
μ_{10}	1.58	0.93	1.58	1.40	1.46	0.13	0.22	0.16	1.50	0.26
μ_{11}	1.58	1.87	1.66	1.70	1.65	0.13	0.17	0.14	0.16	0.14
δ_0	0.61	0.54	1.19	0.93	1.18	0.20	0.36	0.33	1.61	0.63
δ_1	0.61	0.69	0.62	0.49	0.56	0.20	0.29	0.24	0.37	0.29
δ_0	0.00	0.79	0.65	0.73	0.81	0.00	0.36	0.36	0.50	0.60
δ_1	0.00	0.94	0.08	0.30	0.19	0.00	0.28	0.16	1.50	0.26
τ	0.00	0.15	-0.57	-0.43	-0.62	0.00	0.45	0.37	1.59	0.67
Assuming $\tau = 0$										
μ_{00}	0.97	0.37	0.73	0.51	0.61	0.16	0.22	0.20	0.27	0.26
μ_{01}	0.97	1.22	1.07	1.14	0.91	0.16	0.22	0.17	0.32	0.22
μ_{10}	1.58	0.97	1.32	1.11	1.42	0.12	0.21	0.18	0.38	0.18
μ_{11}	1.58	1.83	1.66	1.74	1.72	0.12	0.16	0.13	0.17	0.14
$\delta_0 = \delta_1$	0.61	0.61	0.59	0.60	0.80	0.20	0.24	0.20	0.37	0.26
$\delta_0 = \delta_1$	0.00	0.86	0.34	0.62	0.30	0.00	0.24	0.18	0.35	0.17

Table 2

Simulation results for correct models: bias and standard deviation in estimating the μ_s and the derived quantities using the unadjusted (UA), PA, OR, IPW and AIPW methods, with and without allowing a nonzero interaction between t and s (i.e., $\tau = 0$). Outcome and treatment data are simulated for the same 128 U.S. patients in the quality of life example of Section 4 or a larger set of 640 patients with the same characteristics. Each entry is based on 10^4 replicates

Parameter	Bias					Standard deviation				
	UA	PA	OR	IPW	AIPW	UA	PA	OR	IPW	AIPW
$n = 128$, not assuming $\tau = 0$										
μ_{00}	0.57	-0.19	0.00	-0.02	0.00	0.11	0.34	0.25	0.61	0.42
μ_{01}	-0.08	0.07	0.00	-0.02	0.00	0.11	0.19	0.16	0.31	0.21
μ_{10}	0.00	-0.59	0.00	-0.05	0.00	0.07	0.30	0.22	1.81	0.48
μ_{11}	-0.08	0.13	0.00	0.00	0.00	0.07	0.13	0.10	0.13	0.11
δ_0	-0.56	-0.41	0.00	-0.03	0.00	0.13	0.46	0.32	1.95	0.64
δ_1	0.00	0.05	0.00	0.02	0.00	0.13	0.23	0.18	0.34	0.24
δ_0	-0.65	0.26	0.00	0.01	0.00	0.00	0.42	0.30	0.67	0.47
δ_1	-0.08	0.72	0.00	0.05	0.00	0.00	0.35	0.24	1.81	0.50
τ	0.57	0.46	0.00	0.04	0.00	0.00	0.54	0.37	1.97	0.68
$n = 128$, assuming $\tau = 0$										
μ_{00}	0.23	-0.31	0.00	-0.20	0.00	0.10	0.25	0.18	0.33	0.23
μ_{01}	-0.11	0.18	0.00	0.04	0.00	0.10	0.20	0.12	0.27	0.17
μ_{10}	0.26	-0.44	0.00	-0.27	0.00	0.08	0.27	0.18	0.40	0.23
μ_{11}	-0.07	0.05	0.00	-0.03	0.00	0.08	0.12	0.09	0.15	0.11
$\delta_0 = \delta_1$	0.03	-0.13	0.00	-0.07	0.00	0.12	0.23	0.13	0.33	0.21
$\delta_0 = \delta_1$	-0.34	0.49	0.00	0.24	0.00	0.00	0.29	0.19	0.37	0.23
$n = 640$, not assuming $\tau = 0$										
μ_{00}	0.57	-0.18	0.00	0.00	0.00	0.05	0.14	0.10	0.19	0.14
μ_{01}	-0.08	0.07	0.00	0.00	0.00	0.05	0.08	0.07	0.11	0.08
μ_{10}	0.00	-0.59	0.00	-0.01	0.00	0.03	0.13	0.09	0.47	0.16
μ_{11}	-0.08	0.13	0.00	0.00	0.00	0.03	0.06	0.04	0.05	0.04

Parameter	Bias					Standard deviation				
	UA	PA	OR	IPW	AIPW	UA	PA	OR	IPW	AIPW
δ_0	-0.57	-0.40	0.00	0.00	0.00	0.06	0.20	0.14	0.53	0.21
δ_1	0.00	0.05	0.00	0.00	0.00	0.06	0.10	0.08	0.12	0.09
δ_0	-0.65	0.26	0.00	0.00	0.00	0.00	0.18	0.13	0.21	0.16
δ_1	-0.08	0.72	0.00	0.01	0.00	0.00	0.15	0.10	0.47	0.16
τ	0.57	0.46	0.00	0.00	0.00	0.00	0.23	0.16	0.54	0.23
$n = 640$, assuming $\tau = 0$										
μ_{00}	0.24	-0.30	0.00	-0.04	0.00	0.05	0.11	0.08	0.17	0.11
μ_{01}	-0.10	0.18	0.00	0.01	0.00	0.05	0.09	0.05	0.11	0.08
μ_{10}	0.26	-0.43	0.00	-0.05	0.00	0.03	0.12	0.07	0.19	0.11
μ_{11}	-0.07	0.05	0.00	0.00	0.00	0.03	0.05	0.04	0.06	0.05
$\delta_0 = \delta_1$	0.03	-0.13	0.00	-0.02	0.00	0.06	0.10	0.06	0.13	0.09
$\delta_0 = \delta_1$	-0.34	0.48	0.00	0.05	0.00	0.00	0.12	0.08	0.18	0.11

Table 3

Simulation results for misspecified models: bias and standard deviation in estimating the μ_{ts} and the derived quantities using the unadjusted (UA), PA, OR, IPW, and AIPW methods, with misspecified OR and/or PS models. Outcome and treatment data are simulated for the same 128 U.S. patients in the quality of life example of Section 4. Each entry is based on 10^4 replicates

Parameter	Bias				Standard deviation					
	UA	PA	OR	IPW	AIPW	UA	PA	OR	IPW	AIPW
OR model incorrect, PS model correct										
μ_{00}	0.57	-0.19	0.14	-0.03	0.02	0.11	0.35	0.29	0.62	0.44
μ_{01}	-0.08	0.07	-0.08	-0.03	-0.02	0.11	0.19	0.17	0.28	0.23
μ_{10}	0.00	-0.59	-0.23	-0.06	-0.05	0.07	0.31	0.29	1.56	0.73
μ_{11}	-0.08	0.13	0.05	0.00	0.00	0.07	0.12	0.12	0.13	0.13
δ_0	-0.57	-0.40	-0.37	-0.03	-0.06	0.13	0.47	0.40	1.73	0.86
δ_1	0.00	0.06	0.13	0.03	0.02	0.13	0.23	0.20	0.32	0.27
δ_0	-0.65	0.26	-0.22	0.01	-0.04	0.00	0.42	0.34	0.67	0.50
δ_1	-0.08	0.72	0.28	0.06	0.05	0.00	0.35	0.32	1.56	0.73
τ	0.57	0.46	0.50	0.06	0.09	0.00	0.55	0.46	1.75	0.89
OR model correct, PS model incorrect										
μ_{00}	0.57	-0.19	0.00	0.22	0.00	0.11	0.34	0.25	0.79	0.40
μ_{01}	-0.08	0.07	0.00	-0.07	0.00	0.11	0.19	0.16	0.22	0.18
μ_{10}	0.00	-0.59	0.00	-0.38	0.00	0.07	0.31	0.21	0.87	0.33
μ_{11}	-0.08	0.13	0.00	0.07	0.00	0.07	0.13	0.10	0.14	0.10
δ_0	-0.56	-0.40	0.00	-0.60	0.00	0.13	0.46	0.32	1.25	0.52
δ_1	0.00	0.06	0.00	0.14	0.00	0.13	0.23	0.18	0.27	0.21
δ_0	-0.65	0.27	0.00	-0.29	0.00	0.00	0.41	0.30	0.82	0.45
δ_1	-0.08	0.72	0.00	0.44	-0.01	0.00	0.35	0.24	0.89	0.35
τ	0.57	0.45	0.00	0.74	0.00	0.00	0.54	0.37	1.29	0.56
both models incorrect										
μ_{00}	0.57	-0.19	0.14	0.21	0.12	0.10	0.34	0.28	0.69	0.45
μ_{01}	-0.08	0.07	-0.08	-0.07	-0.05	0.10	0.19	0.17	0.23	0.20

Parameter	Bias				Standard deviation					
	UA	PA	OR	IPW	AIPW	UA	PA	OR	IPW	AIPW
μ_{10}	0.00	-0.59	-0.23	-0.39	-0.24	0.07	0.31	0.29	0.77	0.43
μ_{11}	-0.08	0.13	0.05	0.07	0.06	0.07	0.12	0.12	0.14	0.13
δ_0	-0.57	-0.40	-0.36	-0.60	-0.36	0.13	0.46	0.40	1.12	0.63
δ_1	0.00	0.06	0.13	0.14	0.11	0.13	0.23	0.20	0.27	0.24
δ_0	-0.65	0.26	-0.21	-0.28	-0.17	0.00	0.41	0.34	0.73	0.50
δ_1	-0.08	0.72	0.28	0.46	0.30	0.00	0.35	0.32	0.79	0.46
τ	0.57	0.46	0.49	0.74	0.47	0.00	0.54	0.46	1.16	0.68