Published in final edited form as:

J Am Stat Assoc. 2016 ; 111(514): 538–548. doi:10.1080/01621459.2015.1130633.

Joint Estimation of Treatment and Placebo Effects in Clinical Trials with Longitudinal Blinding Assessments

Wei Liu^{1,2}, Zhiwei Zhang^{2,*}, R. Jason Schroeder², Martin Ho², Bo Zhang², Cynthia Long³, Hui Zhang⁴, and Telba Z. Irony⁵

¹Department of Mathematics, Harbin Institute of Technology, Harbin, P. R. China

²Division of Biostatistics, Office of Surveillance and Biometrics, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland, USA

³Division of Reproductive, Gastro-Renal, and Urological Devices, Office of Device Evaluation, Center for Devices and Radiological Health, Food and Drug Administration, Silver Spring, Maryland, USA

⁴Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee, USA

⁵Office of Biostatistics and Epidemiology, Center for Biologic Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland, USA

Abstract

In some therapeutic areas, treatment evaluation is frequently complicated by a possible placebo effect (i.e., the psychobiological effect of a patient's knowledge or belief of being treated). When a substantial placebo effect is likely to exist, it is important to distinguish the treatment and placebo effects in quantifying the clinical benefit of a new treatment. These causal effects can be formally defined in a joint causal model that includes treatment (e.g., new versus placebo) and treatmentality (i.e., a patient's belief or mentality about which treatment she or he has received) as separate exposures. Information about the treatmentality exposure can be obtained from blinding assessments, which are increasingly common in clinical trials where blinding success is in question. Assuming that treatmentality has a lagged effect and is measured at multiple time points, this article is concerned with joint evaluation of treatment and placebo effects in clinical trials with longitudinal follow-up, possibly with monotone missing data. We describe and discuss several methods adapted from the longitudinal causal inference literature, apply them to a weight loss study, and compare them in simulation experiments that mimic the weight loss study.

Keywords

causal inference; confounding; double robustness; G-computation; inverse probability weighting; sequential regression; targeted minimum loss based estimation; treatmentality

* zhiwei.zhang@fda.hhs.gov.

1 Introduction

In some therapeutic areas, treatment evaluation is frequently complicated by a potential placebo effect, i.e., the psychobiological effect of a patient's knowledge or belief of being treated (e.g., Beecher, 1955; Finniss et al., 2010; Hrobjartsson, Kaptchuk and Miller, 2011). Here we give the term "placebo" a broad interpretation which includes an inert pill, an inactive device and a sham procedure. A placebo effect is often suspected for a relatively subjective endpoint such as pain, depression, or quality of life. Even for an objective endpoint such as weight loss, a placebo can have a substantial effect through behavioral changes (e.g., diet and exercise). In clinical studies where a placebo effect is plausible, it is important to distinguish the placebo and treatment-specific effects in quantifying the clinical benefit of a new treatment. The standard approach to this problem is a randomized, double-blind, placebo-controlled study. A simple comparison of treatment groups in such a study is thought to represent the causal effect of the experimental treatment, separate from any placebo effect. However, blinding procedures are not equally effective, and even effective blinding does not guarantee that all patients have the same belief or mentality about which treatment they have received (henceforth abbreviated as treatmentality) at all times. A patient who is initially blinded to his or her treatment assignment may over time develop an opinion on the treatment received, based on his or her post-treatment 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, raising concerns that the observed treatment difference may be due in part to differential distributions of treatmentality.

This research was motivated by a randomized, double-blind, placebo-controlled study of an investigational device for weight reduction in obese adult patients. The device was an implantable neuromodulation system designed to deliver intermittent, programmed electrical charge to the intra-abdominal vagus nerve with the intent of reducing hunger and increasing satiety. The placebo control in this case was the same device in an inactive mode (unable to deliver charge to the vagus nerve because the leads were not connected). The same surgical procedure (with anesthesia and the same number of incisions) was used to implant the active device and the placebo control. Subjects and assessors (who performed follow-up evaluations) were blinded to the actual treatment through 12 months of follow-up. Table 1 summarizes the availability of subjects at the month-6 and month-12 visits. The primary endpoint is the percent excess weight loss (%EWL; based on an ideal weight determined by a subject's height and an ideal BMI of 25 kg/m^2) at 12 months, and the relevant data are summarized in Table 2. The average %EWL at 12 months is 36.3 (sd = 28.5) in the active treatment group and 22.7 (sd = 24.6) in the placebo group, with a difference of 13.6 (95% CI: 7.7 to 19.4), which is significantly greater than 0 at the usual level (0.025, one-sided). However, questions remain as to how much of the observed treatment difference is attributable to a placebo effect. Although the belief of being treated does not change a patient's weight immediately, it could instill a positive attitude to life, induce positive changes in diet and exercise, and eventually lead to weight reduction. This study included a blinding assessment component in which patients were asked to guess which treatment they believed they had received, with the option of answering "I don't know" if they did not have

a strong belief, at 1 week (shortly after implantation), 6 months and 12 months. The results, shown in Table 3, suggest that the two groups initially had similar distributions of treatmentality, which over time shifted toward the true treatment in each group. At 12 months, the majority of subjects were able to make correct guesses in both groups. Subjects may have been unblinded because they felt the electrical charge of the active device, or because of adverse events or notable increases in satiety. Some of the possible reasons for unblinding are acceptable or even desirable; thus, partial unblinding alone does not necessarily invalidate a study. The focus of this article is not on measuring blinding success but on accounting for partial unblinding in treatment evaluation.

It is important to recognize that treatmentality is a post-randomization variable and therefore should not be treated as a baseline confounder. It was suggested recently that treatmentality should be treated as another exposure (in addition to treatment) in a joint causal analysis of treatment and placebo effects (Jamshidian, Hubbard and Jewell, 2011; Roy, 2012; Zhang et al., 2013). These authors took different approaches to measuring treatmentality. Jamshidian, Hubbard and Jewell (2011) used treatment-related adverse events as a surrogate for treatmentality, based on clinical considerations in a specific study. Roy (2012) proposed to measure treatmentality by the randomization ratio (active treatment to placebo), which can be varied within a study to allow estimation of the placebo effect. This design, termed “randomized treatment-belief trials”, effectively eliminates confounding for treatmentality through randomization. On the other hand, the randomization ratio as a surrogate for treatmentality does not account for the variability of treatmentality between subjects and over time. Zhang et al. (2013) proposed to measure treatmentality using blinding assessment data (such as those in the weight loss study described earlier). Blinding assessment has become increasingly common in clinical studies where blinding success is in question (e.g., Fuller et al., 1984; Turner et al., 2002; Brinjikji et al., 2010; Freed et al., 2014), and several measures of blinding success have been developed (James et al., 1996; Bang, Ni and Davis, 2004; Bang et al., 2010). Zhang et al. (2013) pointed out that such data can also be useful in joint estimation of treatment and placebo effects. These references (Jamshidian, Hubbard and Jewell, 2011; Roy, 2012; Zhang et al., 2013) assume that treatmentality is a point (or time-independent) exposure. This is clearly not the case in the weight loss study, where treatmentality varied over time and likely had a lagged effect.

This article gives an overview of statistical methods that account for the time-dependent nature of treatmentality in the weight loss study and that take advantage of the longitudinal blinding assessment data available. This problem has been considered by Hubbard, Jamshidian and Jewell (2012), who described and illustrated a G-computation (GC) approach (Robins, 1986) and a targeted minimum loss based estimation (TMLE) approach (van der Laan and Gruber, 2011; van der Laan and Rose, 2011) in the context of a diabetes study (Backonia et al., 1998) with treatment-related adverse events as a surrogate for treatmentality. In this article, we also consider a sequential regression (SR) approach (Robins, 2000), an inverse probability weighting (IPW) approach (Horvitz and Thompson, 1952; Robins, Rotnitzky and Zhao, 1994; Robins, Hernan and Brumback, 2000), and an augmented IPW (AIPW) approach (Scharfstein, Rotnitzky and Robins, 1999; van der Laan and Robins, 2003; Bang and Robins, 2005). All of these approaches are able to adjust for time-dependent confounders for treatmentality under the assumption of no unmeasured

confounders, and to deal with monotone missing data under a sequential ignorability assumption. In addition to these general assumptions, the various methods also require different modeling assumptions and have different theoretical properties and operating characteristics. In particular, the GC and SR estimators are most efficient under correctly specified models, and the AIPW and TMLE estimators are doubly robust and locally efficient (van der Laan and Robins, 2003; van der Laan and Rose, 2011). We describe, discuss and compare these methods in the context of the weight loss study, with close attention to special issues concerning blinding assessment data.

In the next section, we set up the notation and state the key assumptions. We then describe and discuss the various methods in Section 3. The methods are applied to the weight loss study in Section 4 and evaluated in simulation experiments in Section 5. The article ends with a discussion in Section 6.

2 Notation and Assumptions

Consider a randomized clinical trial comparing an experimental treatment ($T=1$) with placebo ($T=0$). Suppose patients are evaluated at baseline and K follow-up visits. Let S_k denote a patient's treatment at time $k=0, \dots, K-1$, where $k=0$ corresponds to baseline. Following Zhang et al. (2013), we assume that the sample space $\mathcal{S}$ for each S_k has 3 points: 1 (active treatment), 0 (placebo), and u (unknown). Let L_k be a set of baseline ($k=0$) and intermediate ($k=1, \dots, K-1$) confounders for S_k , and let $Y=L_K$ denote the outcome of primary interest, to be measured at time K . In the weight loss study, $K=2$, $Y=L_2$ is %EWL at 12 months, L_1 is %EWL at 6 months, and L_0 consists of age (in years), gender, race (white or non-white), diabetes (presence or absence of Type 2 diabetes mellitus), body mass index (BMI; in kg/m^2), hunger score (a subscale of the Three-Factor Eating Questionnaire), and depression score (based on Beck Depression Inventory-II) at baseline. Following standard notation in causal inference, we write

$$\begin{aligned} A_0 &= (T, S_0), & A_k &= S_k, & k &= 1, \dots, K-1; \\ \bar{S}_k &= (S_0, \dots, S_k), & \bar{A}_k &= (A_0, \dots, A_k) = (T, \bar{S}_k), & k &= 0, \dots, K-1; \\ \bar{L}_k &= (L_0, \dots, L_k), & k &= 0, \dots, K; \\ \bar{S} &= \bar{S}_{K-1}, & \bar{A} &= \bar{A}_{K-1}, & \bar{L} &= \bar{L}_K. \end{aligned}$$

Similar notation will be used throughout without further comment.

Our causal questions can be formulated in terms of potential outcomes (Rubin, 1974). For each exposure history $\mathbf{a} = (a_1, \dots, a_{K-1}) = (t, \bar{s})$ with $t \in \{0, 1\}$ and $\bar{s} \in \mathcal{S}^K$, we write $\bar{L}(\mathbf{a})$ for the response history that will result if the subject is exposed according to $\mathbf{a}$, and $\mu(\mathbf{a}) = E\{Y(\mathbf{a})\} = E\{L_K(\mathbf{a})\}$ for the mean outcome at time K . For each $\bar{s}$, a treatment effect can be defined as $\mathcal{E}(\cdot, \bar{s}) = \mu(1, \bar{s}) - \mu(0, \bar{s})$, which is the change in the mean outcome that would result if we switched treatment from 0 to 1 while fixing treatment at $\bar{s}$ for the entire population. Such treatment effects may be relevant to consider in regulatory settings, despite the hypothetical nature of this definition. Similarly, a placebo effect may be defined at each treatment level t as $\mathcal{E}(t, \cdot) = \mu(t, \bar{1}) - \mu(t, \bar{0})$, where $\bar{1}$ and $\bar{0}$ denote vectors of 1's and 0's,

respectively, of suitable lengths (implied by the context). Thus, $\delta(t, \cdot)$ is a comparison of the hypothetical scenarios where all patients receive treatment t and believe, throughout the study, that they have received treatment 1 (or 0). Although this definition of placebo effects is focused on comparing $\bar{s} = \bar{1}$ with $\bar{s} = \bar{0}$, it is certainly possible to assess the effect of treatmentality through other comparisons. The treatment and placebo effects defined above can be regarded as controlled direct and indirect effects if treatmentality is considered a mediator (Robins and Greenland, 1992; Zhang et al., 2013). In addition, one may be interested in a possible interaction between treatment and treatmentality, defined as $\tau = \delta(\cdot, \bar{1}) - \delta(\cdot, \bar{0}) = \delta(1, \cdot) - \delta(0, \cdot)$. This article is concerned with the general problem of estimating the means $\mu(\cdot)$ and the related causal effects and interaction.

The potential outcomes will be related to the actual outcomes through the assumption of consistency:

$$\bar{L}_k = \bar{L}_k(\bar{a}) \quad \text{if} \quad \bar{A}_{k-1} = \bar{a}_{k-1}.$$

This implies that causes precede effects in time: if two exposure histories are identical up to time $k-1$, then the corresponding potential outcomes are identical up to time k . Now we can write

$$E(Y|T=t) = \sum_{\bar{s} \in \mathcal{S}^K} P(\bar{S}=\bar{s}|T=t) E\{Y(t, \bar{s})|T=t, \bar{S}=\bar{s}\},$$

which is difficult to simplify if a placebo effect exists at treatment level t (i.e., if $Y(t, \bar{s})$ depends on $\bar{s}$). Consequently, the intent-to-treat analysis with focus on $E(Y|T=1) - E(Y|T=0)$ does not generally have a clear interpretation in the present causal framework. One may be tempted to estimate $\mu(\cdot) = \mu(t, \bar{s})$ as $E(Y|T=t, \bar{S}=\bar{s})$; however, this approach requires a strong assumption:

$$P\{\bar{S}=\bar{s}|T=t, Y(t, \bar{s})\} = P(\bar{S}=\bar{s}|T=t) \quad \text{with probability 1.}$$

This assumption is easily violated in the presence of confounders (i.e., baseline and post-treatment variables that predict both treatmentality and the outcome of interest).

In this article, we allow there to be (possibly time-dependent) confounders but assume (for the most part) that all confounders be measured in the sense that

$$A_k \perp \bar{L}(\bar{a}) | \bar{L}_k, \bar{A}_{k-1} = \bar{a}_{k-1}, \quad k=0, \dots, K-1, \quad (1)$$

where $\perp$ denotes conditional independence, and we set $\bar{a}_{-1} = \bar{a}_0 = 0$ for convenience. This assumption generalizes the strongly ignorable treatment assignment assumption for a point treatment (Rosenbaum and Rubin, 1983). Because T is randomized and therefore independent of all baseline variables, the essence of assumption (1) is the conditional

independence of S_k and $L(\cdot)$ given the response history up to time k and the exposure history up to time $k-1$. In other words, given the observed past, we assume that S_k is assigned randomly as in a sequential randomized trial. Furthermore, nonparametric identification of $\mu(\cdot)$ requires to satisfy a positivity assumption:

$$f(A_k=a_k|\bar{A}_{k-1}=\bar{a}_{k-1}, \bar{L}_k) > 0 \quad \text{whenever} \quad f(\bar{A}_{k-1}=\bar{a}_{k-1}, \bar{L}_k) > 0, \quad (2)$$

where f denotes a generic (conditional) density with respect to a suitable measure. Because T is randomized and S is discrete, the real requirement on $\pi = (t, \bar{s})$ is that $P(S_k = s_k | T = t, \bar{S}_{k-1} = \bar{s}_{k-1}, \bar{L}_k) > 0$ as long as the conditioning event is a realistic configuration. Together, assumptions (1) and (2) ensure that $\mu(\cdot)$ is identified nonparametrically, provided $(\cdot, \bar{L})$ is fully observed (e.g., van der Laan and Robins, 2003).

Missing data arise frequently in clinical trials because subjects may drop out of a study for various reasons (e.g., adverse events, lack of improvement, or simply loss of interest). In the weight loss study, for example, about 14% of subjects did not make the month-12 visit (see Table 1). Let $R_k = 1$ if a subject is evaluated at time k , 0 otherwise. We shall focus on monotone missingness in the sense that

$$1 \equiv R_0 \geq R_1 \geq \dots \geq R_K, \quad (3)$$

which is reasonable to expect in clinical trials where patient dropout is the primary reason for missingness (Molenberghs and Kenward, 2007). This assumption is met for about 97% of the subjects in the weight loss study. We assume that missing data are sequentially ignorable in the sense that

$$P(R_k=1|\bar{A}, \bar{L}, R_{k-1}=1) = P(R_k=1|\bar{A}_{k-1}, \bar{L}_{k-1}, R_{k-1}=1), \quad k=1, \dots, K. \quad (4)$$

This is just a longitudinal version of the general missing-at-random assumption (Rubin, 1976). Finally, we make another positivity assumption here:

$$P(R_k=1|\bar{A}_{k-1}, \bar{L}_{k-1}, R_{k-1}=1) > 0 \text{ with probability } 1, \quad k=1, \dots, K. \quad (5)$$

Assumptions (4) and (5) together imply that the joint distribution of $(\cdot, \bar{L})$ is identified nonparametrically from the observable $(C, C \wedge (K-1), \bar{L}_C)$, where $C = \max\{k : R_k = 1\}$ is the censoring time (i.e., the last time at which a subject is evaluated) and $\wedge$ denotes minimum. This further implies the identifiability of $\mu(\cdot)$ for any π that satisfies assumptions (1) and (2).

3 Methods

In this section, we describe and discuss estimators of $\mu(\cdot)$ based on a random sample of size n from the joint law of $(C, C_{\Lambda(K-1)}, \bar{L}_C)$, under assumptions (1)–(5) and additional modeling assumptions. Our discussion of specific estimators will be focused on consistency or asymptotic unbiasedness, as asymptotic normality is usually immediate from standard M -estimation theory (e.g., van der Vaart, 1998; Stefanski and Boos, 2002). Asymptotic variance formulas are straightforward to derive but cumbersome to present and implement in practice. For ease of implementation, we use nonparametric bootstrap standard errors and confidence intervals for inference.

3.1 G-computation (GC)

The GC approach (Robins, 1986) is based on the following factorization for the joint density of $\bar{L}(\cdot)$:

$$f(\bar{L}(\bar{a})) = f(L_0(\bar{a})) \prod_{k=1}^K f(L_k(\bar{a}) | \bar{L}_{k-1}(\bar{a})),$$

which implies that

$$\mu(\bar{a}) = \int f(L_0(\bar{a})) \left\{ \prod_{k=1}^{K-1} f(L_k(\bar{a}) | \bar{L}_{k-1}(\bar{a})) \right\} E\{Y(\bar{a}) | \bar{L}_{K-1}(\bar{a})\} d\bar{L}_{K-1}(\bar{a}), \quad (6)$$

where the integral is with respect to a suitable measure (suppressed from the notation). Because $L_0(\cdot) \equiv L_0$ is always observed, its distribution can be estimated by the empirical distribution of L_0 . For $k > 0$, repeated applications of assumption (1) yield

$$f(L_k(\bar{a}) | \bar{L}_{k-1}(\bar{a}) = \bar{l}_{k-1}) = f(L_k = l_k | \bar{L}_{k-1} = \bar{l}_{k-1}, \bar{A}_{k-1} = \bar{a}_{k-1}), \quad (7)$$

which helps relate the potential outcome distributions to the actual outcome distributions. Assumption (2) further implies that the right-hand side of (7) is empirically identifiable without missing data. In the presence of missing data, assumption (4) allows us to write

$$f(L_k = l_k | \bar{L}_{k-1} = \bar{l}_{k-1}, \bar{A}_{k-1} = \bar{a}_{k-1}) = f(L_k = l_k | \bar{L}_{k-1} = \bar{l}_{k-1}, \bar{A}_{k-1} = \bar{a}_{k-1}, R_k = 1), \quad (8)$$

and assumption (5) implies that the right-hand side of (8) is identifiable from the observed data.

The foregoing discussion suggests that parametric models, say $E(Y | \bar{L}_{K-1}, \bar{A}_{K-1}, R_K = 1) = m_K(\bar{L}_{K-1}, \bar{A}_{K-1}; \alpha_K)$ and $f(L_k | \bar{L}_{k-1}, \bar{A}_{k-1}, R_k = 1; \alpha_k)$, $k = 1, \dots, K-1$, be specified for the

empirical quantities that are involved in (6) through (7) and (8). These models may be specified as (generalized) linear models, and the α_k may be estimated from the observed data by (iteratively reweighted) least squares. Let $\hat{\alpha}_k$ denote the resulting estimate of α_k , $k = 1, \dots, K$. Then expression (6) suggests that $\mu(\cdot)$ may be estimated by

$$\hat{\mu}_{gc}(\bar{a}) = \int \hat{f}(L_0) \left\{ \prod_{k=1}^{K-1} f(L_k | \bar{L}_{k-1}, \bar{A}_{k-1} = \bar{a}_{k-1}, R_k = 1; \hat{\alpha}_k) \right\} m_K(\bar{L}_{K-1}, \bar{A} = \bar{a}; \hat{\alpha}_K) d\bar{L}_{K-1},$$

where $\hat{f}(L_0)$ denotes the empirical distribution of L_0 .

In the context of the weight loss study ($K = 2$), this approach generally requires two models: a parametric mean model $E(Y | \bar{L}_1, R_2 = 1) = m_2(\bar{L}_1; \alpha_2)$ and a parametric distribution model $f(L_1 | L_0, A_0, R_1 = 1; \alpha_1)$. We work with a linear m_2 -model with possible covariate configurations given in Web Table 1 (see Web Appendix A for details of model selection). Note that this study has a one-dimensional continuous L_1 , for which a (possibly transformed) linear model with normality may be appropriate. Because $m_2(\bar{L}_1; \alpha_2)$ happens to be linear in L_1 in Web Table 1, the model for L_1 can be reduced to a parametric mean model $E(L_1 | L_0, A_0, R_1 = 1; \alpha_1) = m_1(L_0, A_0; \alpha_1)$, which is also specified as a linear model with possible covariate configurations given in Web Table 1. Now the estimator $\hat{\mu}_{gc}(\cdot)$ reduces to $\hat{E}\{m_2(L_0, m_1(L_0, A_0 = a_0; \hat{\alpha}_1), \bar{A} = \bar{a}; \hat{\alpha}_2)\}$, where $\hat{E}$ denotes sample average (with respect to L_0 in this case).

3.2 Sequential regression (SR)

The SR approach (Robins, 2000) is similar in spirit to GC but is based on a different representation:

$$\begin{aligned} \mu(\bar{a}) &= E\{Y(\bar{a})\} \\ &= E[E\{Y(\bar{a}) | L_0\}] \\ &= E[E[E\{Y(\bar{a}) | \bar{L}_1(\bar{a})\} | L_0]] \\ &= E[E[\dots E[E\{Y(\bar{a}) | \bar{L}_{K-1}(\bar{a})\} | \bar{L}_{K-2}(\bar{a})] \dots | L_0]], \quad (9) \end{aligned}$$

which follows from the law of iterated expectations. Expressions (7) and (8), with $k = K$, imply that

$$E\{Y(\bar{a}) | \bar{L}_{K-1}(\bar{a}) = \bar{l}_{K-1}\} = E(Y | \bar{L}_{K-1} = \bar{l}_{K-1}, \bar{A} = \bar{a}, R_K = 1) =: h_{K-1}(\bar{l}_{K-1}). \quad (10)$$

Setting $k = K - 1$ in (7) and (8), with $L_{K-1}(\cdot)$ replaced by $h_{K-1}(\bar{L}_{K-1}(\cdot))$, yields

$$E\{h_{K-1}(\bar{L}_{K-1}(\bar{a}))|\bar{L}_{K-2}(\bar{a})=\bar{l}_{K-2}\}=E\{h_{K-1}(\bar{L}_{K-1})|\bar{L}_{K-2}=\bar{l}_{K-2}, \bar{A}_{K-2}=\bar{a}_{K-2}, R_{K-1}=1\}=h_{K-2}(\bar{l}_{K-2}).$$

$$(11)$$

Proceed in this manner, with decreasing k , to define $h_{K-3}(l_{K-3}), \dots, h_0(l_0)$; then expression (9) implies that $\mu(\cdot) = E\{h_0(L_0)\}$.

Motivated by the above observations, the SR approach starts with a regression analysis of Y on $(\bar{L}_{K-1}, \bar{A}_{K-1})$ among subjects with complete data ($R_K = 1$). Let $\hat{h}_{K-1}$ denote the resulting estimate of h_{K-1} according to (10). Next, regress $\hat{h}_{K-1}(\bar{L}_{K-1})$ on $(\bar{L}_{K-2}, \bar{A}_{K-2})$ among subjects with $R_{K-1} = 1$, and obtain $\hat{h}_{K-2}$, an estimate of h_{K-2} based on (11). Repeat this procedure with decreasing k to obtain $\hat{h}_0$, an estimate of h_0 . The SR estimator of $\mu(\cdot)$ is then given by $\hat{\mu}_{\text{sr}}(\cdot) = \hat{E}\{\hat{h}_0(L_0)\}$.

The SR approach requires a series of parametric mean models but no parametric distribution models. In contrast, the GC approach generally requires a series of parametric distribution models, $f(L_k|\bar{L}_{k-1}, \bar{A}_{k-1}, R_k = 1; \alpha_k)$, $k = 1, \dots, K-1$, in addition to a parametric mean model, $m_K(\bar{L}_{K-1}, \bar{A}_{K-1}; \alpha_K)$. Because the models for GC are sufficient to induce the models for SR, the latter approach is generally more robust than the former approach in terms of modeling assumptions. On the other hand, the models involved in GC may be easier to interpret and thus better able to incorporate subject matter knowledge.

In the weight loss study, the SR approach requires two parametric mean models:

$$\begin{aligned} E(Y|\bar{L}_1, \bar{A}, R_2=1) &= m_2(\bar{L}_1, \bar{A}; \alpha_2), \\ E\{h_1(\bar{L}_1)|L_0, A_0, R_1=1\} &= m_1^\dagger(L_0, A_0; \alpha_1^\dagger). \end{aligned}$$

Because we work with a linear m_2 -model, $h_1(\bar{L}_1)$ is also linear in L_1 , and the second model can be reduced to $E(L_1|L_0, A_0, R_1 = 1) = m_1(L_0, A_0; \alpha_1)$ (introduced in Section 3.1). Thus, in our analysis of the weight loss study data, GC and SR rely on the same models and lead to the same estimator.

3.3 Inverse probability weighting (IPW)

The IPW approach is motivated by the following identity:

$$\begin{aligned} \mu(\bar{a}) &= E \left[\frac{I(\bar{A}=\bar{a}, R_K=1)Y}{P\{\bar{A}=\bar{a}|\bar{L}_{K-1}(\bar{a})\} P(R_K=1|\bar{L}_{K-1}, \bar{A})} \right] \\ &= E \left[\frac{I(\bar{A}=\bar{a}, R_K=1)Y}{\left\{ \prod_{k=0}^{K-1} P(A_k=a_k|\bar{L}_k, \bar{A}_{k-1}=\bar{a}_{k-1}) \right\} \prod_{k=1}^K P(R_k=1|\bar{L}_{k-1}, \bar{A}_{k-1}, R_{k-1}=1)} \right], \end{aligned} \quad (12)$$

which follows from assumptions (1)–(5) and a standard conditioning argument. This representation suggests that parametric models be specified for the conditional probabilities

in the denominator of (12). Because T is randomized, models for μ may be induced from the following models for $\bar{S}$:

$$P(S_k = s_k | \bar{L}_k, T, \bar{S}_{k-1}) = p_k(s_k | \bar{L}_k, T, \bar{S}_{k-1}; \beta_k), \quad k=0, \dots, K-1. \quad (13)$$

Because S_k is categorical with three levels, its distribution could be specified using a two-stage model in which the first stage specifies the conditional probability of $S_k = u$ given the conditioning variables shown above, and the second stage specifies the conditional probability of $S_k = 1$ given $S_k = u$ and the conditioning variables (Zhang et al., 2013, Web Appendix C). Note that assumption (4) implies that

$$P(S_k = s_k | \bar{L}_k, T, \bar{S}_{k-1}) = P(S_k = s_k | \bar{L}_k, T, \bar{S}_{k-1}, R_k = 1), \quad k=1, \dots, K-1,$$

which allows β_k to be estimated from a regression analysis based on the subjects with $R_k = 1$. In addition, the IPW approach requires a series of models for the R_k :

$$P(R_k = 1 | \bar{L}_{k-1}, \bar{A}_{k-1}, R_{k-1} = 1) = q_k(\bar{L}_{k-1}, \bar{A}_{k-1}; \gamma_k), \quad k=1, \dots, K. \quad (14)$$

A standard binary regression model with a logit or probit link may be used for each R_k , and standard software can be used to estimate γ_k from the subjects with $R_{k-1} = 1$. Let $(\hat{\beta}_k, \hat{\gamma}_k)$ denote the estimates of (β, γ) . Then the IPW estimator of $\mu(\bar{a})$ can be obtained as

$$\hat{\mu}_{\text{ipw}}(\bar{a}) = \hat{E} \left[\frac{I(\bar{A} = \bar{a}, R_K = 1)Y}{\hat{\pi}_t \left\{ \prod_{k=0}^{K-1} p_k(s_k | \bar{L}_k, t, \bar{S}_{k-1}; \hat{\beta}_k) \right\} \prod_{k=1}^K q_k(\bar{L}_{k-1}, \bar{a}_{k-1}; \hat{\gamma}_k)} \right],$$

where $\hat{\pi}_t = \hat{E}\{I(T=t)\}$ is the proportion of subjects receiving treatment t . Even though the true value of $\pi_t = P(T=t)$ is known, substituting $\hat{\pi}_t$ for π_t generally leads to better efficiency (Robins, Rotnitzky and Zhao, 1994).

In the weight loss study, we work with a standard logistic regression model for (14) and a two-stage p_k -model (as described above) where each stage is a logistic regression model. Possible covariate configurations in these models are given in Web Table 1.

3.4 Augmented IPW (AIPW)

The AIPW approach is motivated by the form of the efficient influence curve for $\mu(\bar{a})$:

$$h_0(L_0) + \sum_{k=1}^{K-1} \frac{I(\bar{A}_{k-1} = \bar{a}_{k-1}, R_k = 1) \{h_k(\bar{L}_k) - h_{k-1}(\bar{L}_{k-1})\}}{P\{\bar{A}_{k-1} = \bar{a}_{k-1} | \bar{L}_{k-1}(\bar{a})\} P(R_k = 1 | \bar{L}_{k-1}, \bar{A}_{k-1})} + \frac{I(\bar{A} = \bar{a}, R_K = 1) \{Y - h_{K-1}(\bar{L}_{K-1})\}}{P\{\bar{A} = \bar{a} | \bar{L}_{K-1}(\bar{a})\} P(R_K = 1 | \bar{L}_{K-1}, \bar{A})} - \mu(\bar{a}),$$

where the h_k are defined in Section 3.2. A detailed exposition of this theory is given by van der Laan and Robins (2003). Estimating the above expression for each subject and setting the sample average equal to 0, the AIPW estimator of $\mu(\cdot)$ is obtained as

$$\begin{aligned} \hat{\mu}_{\text{aipw}}(\bar{a}) &= \hat{\mu}_{\text{sr}}(\bar{a}) + \hat{E} \left[\sum_{k=1}^{K-1} \frac{I(\bar{A}_{k-1}=\bar{a}_{k-1}, R_k=1) \{ \hat{h}_k(\bar{L}_k) - \hat{h}_{k-1}(\bar{L}_{k-1}) \}}{\hat{\pi}_t \left\{ \prod_{j=0}^{k-1} p_j(s_j | \bar{L}_j, t, \bar{s}_{j-1}; \hat{\beta}_j) \right\} \prod_{j=1}^k q_j(\bar{L}_{j-1}, \bar{a}_{j-1}; \hat{\gamma}_j)} \right. \\ &\quad \left. + \frac{I(\bar{A}=\bar{a}, R_K=1) \{ Y - \hat{h}_{K-1}(\bar{L}_{K-1}) \}}{\hat{\pi}_t \left\{ \prod_{j=0}^{K-1} p_j(s_j | \bar{L}_j, t, \bar{s}_{j-1}; \hat{\beta}_j) \right\} \prod_{j=1}^K q_j(\bar{L}_{j-1}, \bar{a}_{j-1}; \hat{\gamma}_j)} \right] \\ &= \hat{\mu}_{\text{ipw}}(\bar{a}) + \hat{E} \left(\sum_{k=0}^{K-1} \hat{h}_k(\bar{L}_k) \left[\frac{I(\bar{A}_{k-1}=\bar{a}_{k-1}, R_k=1)}{\hat{\pi}_t \left\{ \prod_{j=0}^{k-1} p_j(s_j | \bar{L}_j, t, \bar{s}_{j-1}; \hat{\beta}_j) \right\} \prod_{j=1}^k q_j(\bar{L}_{j-1}, \bar{a}_{j-1}; \hat{\gamma}_j)} \right. \right. \\ &\quad \left. \left. - \frac{I(\bar{A}_k=\bar{a}_k, R_{k+1}=1)}{\hat{\pi}_t \left\{ \prod_{j=0}^k p_j(s_j | \bar{L}_j, t, \bar{s}_j; \hat{\beta}_j) \right\} \prod_{j=1}^{k+1} q_j(\bar{L}_j, \bar{a}_j; \hat{\gamma}_j)} \right] \right), \end{aligned}$$

where an empty product (e.g., $\prod_{j=1}^0$) is defined as 1 for convenience. The estimates involved in this definition (e.g., $\hat{h}_k, \hat{\beta}_j, \hat{\gamma}_j$) are defined in the previous subsections. The first expression for $\hat{\mu}_{\text{aipw}}(\cdot)$ can be interpreted as an augmented SR estimator, with an augmentation term that converges in probability to 0 when the h_k are estimated consistently (even if the p_k - and q_k -models are misspecified). It follows that $\hat{\mu}_{\text{aipw}}(\cdot)$ is consistent as long as the h_k are estimated consistently. The second expression for $\hat{\mu}_{\text{aipw}}(\cdot)$ shows that it augments the IPW estimator, with an augmentation term that converges in probability to 0 under correct p_k - and q_k -models (even if the h_k -models are misspecified). It follows that $\hat{\mu}_{\text{aipw}}(\cdot)$ is also consistent under correct p_k - and q_k -models, and hence doubly robust. If both sets of models are correct, then $\hat{\mu}_{\text{aipw}}(\cdot)$ attains the semiparametric information bound.

3.5 Targeted Minimum Loss Based Estimation (TMLE)

The TMLE approach (van der Laan and Gruber, 2011; van der Laan and Rose, 2011) can be regarded as a modified SR approach that produces a substitution estimator. For a bounded continuous outcome, van der Laan and Gruber (2011) and van der Laan and Rose (2011) recommended the use of a logistic model after re-scaling the outcome to the unit interval. In this section, we assume for simplicity that $K=2$ (as in the weight loss example) and that h_1 and h_0 both follow linear models (as in Web Table 1).

First, fit a linear regression model for $E(Y/\bar{L}_1, \cdot, R_2=1) = m_2(\bar{L}_1, \cdot; a_2)$, which implies that $h_1(\bar{l}_1) = m_2(\bar{l}_1, \cdot; a_2)$. Once a_2 is estimated via least squares, the estimate $\hat{a}_2$ can be substituted to obtain an initial estimate $\hat{h}_1(\bar{l}_1) = m_2(\bar{l}_1, \cdot; \hat{a}_2)$. Define the clever covariate

$$D_2(\bar{a}', r_2, \bar{l}_1) = \frac{I(\bar{a}'=\bar{a}, r_2=1)}{\hat{P}\{\bar{A}=\bar{a} | \bar{L}_1(\bar{a})=\bar{l}_1\} \hat{P}(R_2=1 | \bar{L}_1=\bar{l}_1, \bar{A}=\bar{a})},$$

and fit another linear regression model

$$Y \sim \hat{h}_1(\bar{L}_1) + \varepsilon_2 D_2(\bar{A}, R_2, \bar{L}_1),$$

treating $\hat{h}_1(\bar{L}_1)$ as an offset and ε_2 as the only unknown parameter. This yields

$$\hat{\varepsilon}_2 = \frac{\hat{E}[\{Y - \hat{h}_1(\bar{L}_1)\} D_2(\bar{A}, R_2, \bar{L}_1)]}{\hat{E}\{D_2(\bar{A}, R_2, \bar{L}_1)^2\}},$$

which involves only subjects with $R_1 = 1$ and $R_2 = 1$ because of the indicator in the clever covariate. Now we can update $\hat{h}_1(\bar{L}_1)$ as follows:

$$\hat{h}_1^\bullet(\bar{L}_1) = \hat{h}_1(\bar{L}_1) + \hat{\varepsilon}_2 D_2(\bar{a}, 1, \bar{L}_1).$$

Note that the clever covariate in the above display involves $(\bar{a}, 1)$ instead of $(\bar{a}, R_2)$; in other words, we are making predictions for all subjects with $R_1 = 1$, assuming that they are exposed according to $\bar{a}$ and are followed to the end of the study.

Next, we use $\hat{h}_1^\bullet(\bar{L}_1)$ as the response variable and fit the linear regression model

$$E\{\hat{h}_1^\bullet(\bar{L}_1) | L_0 = a_0, R_1 = 1\} = m_1^\dagger(\bar{L}_0, a_0; \alpha_1^\dagger),$$

where $m_1^\dagger$ is a linear model deduced from m_1 and m_2 . Denote by $\hat{\alpha}_1^\dagger$ the least squares estimate of $\alpha_1^\dagger$, which can be used to define the initial estimate $\hat{h}_0(l_0) = m_1(\bar{l}_0, a_0; \hat{\alpha}_1^\dagger)$. Define another clever covariate

$$D_1(a'_0, r_1, l_0) = \frac{I(a'_0 = a_0, r_1 = 1)}{\hat{P}\{A_0 = a_0 | L_0 = l_0\} \hat{P}\{R_1 = 1 | L_0 = l_0, A_0 = a_0\}},$$

and fit another linear regression model

$$\hat{h}_1^\bullet(\bar{L}_1) \sim \hat{h}_0(L_0) + \varepsilon_1 D_1(A_0, R_1, L_0),$$

treating $\hat{h}_0(L_0)$ as an offset and ε_1 as the only unknown parameter. From this we obtain

$$\hat{\varepsilon}_1 = \frac{E[\{\hat{h}_1^\bullet(\bar{L}_1) - \hat{h}_0(L_0)\} D_1(A_0, R_1, L_0)]}{\hat{E}\{D_1(A_0, R_1, L_0)^2\}},$$

which is based on subjects with $A_0 = a_0$ and $R_1 = 1$. Then we can update $\hat{h}_0(L_0)$ as follows:

$$\hat{h}_0^\bullet(L_0) = \hat{h}_0(L_0) + \hat{\varepsilon}_1 D_1(a_0, 1, L_0).$$

Note that the clever covariate in the above display involves $(a_0, 1)$ instead of (A_0, R_1) , and therefore applies to all subjects. Finally, the TMLE of $\mu(\cdot)$ is given by $\hat{\mu}_{\text{tml}}(\bar{a}) = \hat{E}\hat{h}_0^\bullet(L_0)$.

Like the AIPW estimator, the TML estimator is doubly robust and locally efficient. Unlike the AIPW estimator, the TML estimator is a substitution estimator and therefore respects the natural range of Y . The last property can be important when Y is binary or has well-defined bounds, although this is not the case in the weight loss example.

4 Application

We now analyze the weight loss study data (modified in minor ways to protect confidentiality) using the described methods with the required models, which are specified as generalized linear models with an identity or logit link, depending on the nature of the response variable. Several criteria, including the Bayesian information criterion (BIC), are used to choose covariate configurations in these models (see Web Appendix A and Web Table 1). Because all m_2 -models in Web Table 1 are linear in L_1 , the GC and SR methods are equivalent in this application, and are collectively referred to as GC/SR. We consider two versions of TMLE, implemented using our own R code (TMLE1) and using the `ltmle` package in R (TMLE2). TMLE1 is exactly as described in Section 3.5, and TMLE2 involves a logit link in the h -models as recommended by van der Laan and Gruber (2011) and van der Laan and Rose (2011). For all methods that involve inverse probability weights, we also truncate the weights at 1000, which usually results in a substantial decrease in variance at the expense of a slight increase in bias (results not shown). To comply with the assumption of monotone missingness, we exclude the 15 subjects with intermittent missing data from the main analyses but address intermittent missingness in a subsequent analysis. As a comparator, we also consider a naive method that estimates $\mu(t, \bar{s})$ using $\hat{E}\{I(T=t, \bar{S}=\bar{s}, R_2=1)Y\}/\hat{E}\{I(T=t, \bar{S}=\bar{s}, R_2=1)\}$ without adjusting for confounders and missing data. Throughout this section, we base inference on bootstrap standard errors calculated from 1000 bootstrap samples.

Table 4 shows the point estimates and standard errors for $\mu(t, \bar{s})$ ($t=0, 1$; $\bar{s}=\bar{0}, \bar{1}$) and the associated causal effects (δ 's) and interaction (τ), obtained using the naive method and the GC/SR, IPW, AIPW, TMLE1 and TMLE2 methods. The naive estimates are notably different from the other ones, particular for $\mu(0, \bar{1})$ and the related causal effects. The other five methods, which adjust for confounders and missing data, are largely consistent with each other in terms of point estimates, despite some discrepancies for $\mu(1, \bar{0})$ and the related quantities: $\delta(\cdot, \bar{0})$, $\delta(1, \cdot)$ and τ . The latter observation is apparently due to the small number of subjects (24) in the active treatment group who believed, at both 1 week and 6 months of implantation, that they were in the placebo group. GC/SR and AIPW tend to have smaller standard errors than the other methods. The results in Table 4 are limited to the BIC-based model choices; however, Web Table 2 shows that other model choices lead to similar results. Together, the results in Table 4 and Web Table 2 suggest that both active treatment ($t=1$ vs.

0) and positive treatmentality ($\bar{s} = \bar{1}$ vs. $\bar{0}$) have positive effects on %EWL at 12 months and that they interact with each other in a positive manner, although the level of statistical significance varies across methods and parameters.

The preceding analyses assume that there are no unmeasured confounders and that missing data are sequentially ignorable. Because these assumptions cannot be verified with observed data, it is important to assess the impact of possible violations of these assumptions. In Web Appendix B, we examine the impact of unmeasured confounders using the approaches of Brumback et al. (2004) and Zhang et al. (2014). In Web Appendix C, we consider possible departures from the sequential ignorability assumption under a location shift model. The results of these sensitivity analyses indicate that the estimated causal effects and interaction are fairly robust under mild departures from the assumptions of no unmeasured confounders and sequential ignorability. In addition, Web Appendix D presents a multiple imputation analysis for the 15 subjects with intermittent missing data, and the results are very similar to those in Table 4 and Web Table 2.

The analyses presented in this article should be regarded as exploratory analyses because they were not specified prospectively. Without pre-specification and multiplicity adjustment, their statistical significance is difficult to quantify precisely. Nonetheless, these analyses do yield some interesting findings which can be helpful in regulatory decision making and clinical practice. From a regulatory point of view, an important question is whether active treatment produces any improvement beyond a placebo effect. This question could be formulated in terms of $\delta(\cdot, \bar{1})$ in the counterfactual world where we first make patients believe that they are treated, thereby achieving a placebo effect, and then give active treatment. The relevant results in Table 4, which indicate that such a treatment effect exists and is positive, support regulatory approval of the study device. It is worth noting that a substantial placebo effect may also exist, particularly under active treatment. In addition, there appears to be a positive interaction between active treatment and positive treatmentality, in the sense that their combined effect exceeds the sum of their individual effects (in the absence of the other factor). These observations suggest that clinicians should do everything they can to encourage positive thinking in patients in addition to providing effective treatment. It may be of interest to behavioral scientists to confirm and investigate a positive interaction between active treatment and positive treatmentality.

Indications of a substantial placebo effect and a treatment-treatmentality interaction also highlight the importance of blinding and blinding assessment in future studies of weight loss (and in other therapeutic areas where a non-negligible placebo effect is likely to exist). Without blinding, the observed treatment difference in an intent-to-treat analysis is really an estimate of the average total effect, which is the sum of a treatment effect and a placebo effect (Zhang et al., 2013, page 319). To separate the last two effects, it is important to conceal treatment assignment from patients and assessors and collect blinding assessment data, which can be used to verify blinding success and support a causal analysis. Although a causal analysis is not explicitly concerned with blinding success and does not require perfect blinding, severe unblinding can cause theoretical or practical violations of the positivity assumption (2) for some parameters, leading to identifiability or stability problems (Moore et al., 2012). This is not a serious problem in the weight loss study, but can become one in a

clinical trial with longer follow-up and more frequent blinding assessments. In addition to blinding assessment data, investigators should collect information on possible confounders for treatmentality, which in practice may be taken as important predictors of treatmentality and/or outcome. Frequently, predictors of treatmentality are less well understood than predictors of outcome, and their measurement requires a thoughtful design. In addition to directly measurable quantities such as weight, treatmentality may be related to some psychological factors (e.g., optimism). Our analyses include some psychological factors such as depression and hunger, but it would be helpful to have baseline treatment guess and outcome prediction as measures of optimism (Turner et al., 2002).

5 Simulation

Simulation experiments are performed to evaluate and compare the different methods in situations similar to the weight loss study. We focus on the same set of 464 subjects analyzed in Section 4, with their original values of L_0 , and generate $(T, S_0, L_1, S_1, L_2 = Y, R_1, R_2)$ randomly and sequentially (in the specified order) under the BIC models described in Web Table 1. The parameter values in these models are either estimated from data or modified slightly to create various scenarios with respect to blinding success and treatmentality distribution. Specifically, we consider scenarios with perfect blinding, severe unblinding, more unknowns, and more or less optimism (in terms of guessing active treatment). Here, perfect blinding is defined as no dependence on T in the p_0 -model for $(S_0/L_0, T)$, which forces S_0 to be identically distributed in the two treatment groups but allows S_1 to be differentially distributed due to different distributions of L_1 . This is achieved by centering the terms involving T and setting the corresponding regression coefficients to 0. Severe unblinding is defined as the opposite situation in which the p_0 -model depends heavily on T , and is achieved by doubling the regression coefficients for the centered terms involving T . The other scenarios are created by changing the intercepts in the first or second stage of the p_0 - and p_1 -models in obvious ways. The resulting distributions of S_0 and S_1 for each treatment group are shown in Web Table 9, together with the observed frequency distributions and those corresponding to the estimated BIC models. In each scenario, 10^4 datasets are simulated and analyzed using the BIC models.

Table 5 shows the results of these experiments in terms of empirical bias and standard deviation for estimating $\mu(t, \bar{s})$ ($t = 0, 1$; $\bar{s} = \bar{0}, \bar{1}$) and the associated causal effects ($\mathcal{S}$ s) and interaction (τ). In this context (with %EWL as the outcome), a bias less than 1% would be considered negligible, and a bias greater than 2% could be considered serious. In Table 5, the naive method is clearly biased, and the GC/SR, AIPW and TMLE1 methods are nearly unbiased in all scenarios. The slight bias of the IPW method seems related to its greater variability and is likely a sample size issue, as the method does become less biased with increasing sample size (see Web Appendix E and Web Table 10). The observed bias of the TMLE2 method is apparently due to the inclusion of a logit link in the h -models, which makes the h -models misspecified and represents the main difference between TMLE2 and TMLE1. On the other hand, TMLE2 generally has smaller variability than TMLE1, indicating a bias-variance trade-off between the two versions of TMLE. In terms of variability, the less biased methods (all methods but the naive one) can be ordered roughly as GC/SR < AIPW < TMLE2 < TMLE1 < IPW. Comparing the different scenarios, it seems

that the main issue is whether some treatment-treatment combinations are observed with low frequencies, say $< 15\%$. For example, estimation of $\mu(1, \bar{0})$ and the associated causal effects and interaction is clearly more difficult (in the sense of greater variability and, for IPW, larger bias) in the scenarios of more unknowns and more optimism than in the other scenarios.

We have also performed simulation experiments to examine the consequences of model misspecification (see Web Appendix E for details). This simulation study confirms the double robustness of the AIPW and TMLE methods, although the latter approach appears more sensitive to misspecification of the m -models, which may be reasonable to expect given that the TMLE approach is embedded in the SR framework.

6 Discussion

The methods considered in this article have different theoretical properties and operating characteristics, 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, the AIPW and TMLE methods are worth recommending because of their extra robustness. The GC and SR methods do have an efficiency advantage when the required models are correctly specified, and may be preferable if one is more concerned about efficiency or mean squared error. The IPW method is simple and easy to implement, although it does not appear to outperform the other methods. A simulation study can be helpful in choosing a method for a specific application, and a sensitivity analysis comparing different methods may shed light on the robustness of the results.

The quality of blinding assessment data is essential for the methods described in this article. Here, we have focused on the three-level format for blinding assessment, which is commonly used in practice. With this format, there is a possible misclassification issue in that patients may be tempted to conform to a perceived expectation of remaining blinded and answer “I don’t know” even when they do have a strong belief (Zhang et al., 2013). However, as we explain in Web Appendix C, the methods described in Section 3 remain valid for estimating $\mu(\cdot)$, $\cdot \in \{0, 1\}^{K+1}$, under mild assumptions about the misclassification mechanism. Another concern about the three-level format is that it may be too crude to capture all relevant information about a patient’s treatment. To address this concern, other formats have been proposed that measure treatment in multiple levels (Bang et al., 2010; Brinjikji et al., 2010) or on a continuous scale (Zhang et al., 2013, Section 5). The methods considered here can be adapted to the multi-level format by collapsing the multiple levels into three, but this does not take full advantage of the available information. Methods have yet to be developed for making efficient use of multi-level or continuous treatment data in a causal analysis. A reviewer noted the interesting possibility of collecting validation data for the “unknown” category. How to utilize such validation data in a causal analysis is another open question.

Our consideration of missing data assumes monotone missingness, which is common in clinical trials but not fully general. Although the issue of intermittent missingness is relatively minor in the weight loss study, other trials may suffer from a larger amount of

intermittent missing data. Another issue that often complicates clinical trials is non-compliance, in the sense that patients may fail to comply with the assigned treatment. Furthermore, treatmentality can be associated with missingness, non-compliance and other protocol violations. Further research is needed on how to address these issues appropriately in joint evaluation of treatment and placebo effects.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Backonja M, Beydoun A, Edwards KR, Schwartz SL, Fonseca V, Hes M, LaMoreaux L, Garofalo E. Gabapentin for the symptomatic treatment of painful neuropathy in patients with diabetes mellitus: a randomized controlled trial. *Journal of the American Medical Association*. 1998; 280:1831–1836. [PubMed: 9846777]
- 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]
- Beecher HK. The powerful placebo. *Journal of the American Medical Association*. 1955; 159:1602–1606. [PubMed: 13271123]
- 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]
- Finniss DG, Kaptchuk TJ, Miller F, Benedetti F. Biological, clinical, and ethical advances of placebo effects. *Lancet*. 2010; 375:686–695. [PubMed: 20171404]
- Freed B, Assall OP, Panagiotakis G, Bang H, Park JJ, Moroz A, Baethge C. Assessing blinding in trials of psychiatric disorders: a meta-analysis based on blinding index. *Psychiatry Research*. 2014; 219:241–247. [PubMed: 24930582]
- 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.
- Hrobjartsson A, Kaptchuk TJ, Miller FG. Placebo effect studies are susceptible to response bias and to other types of biases. *Journal of Clinical Epidemiology*. 2011; 64:1223–1229. [PubMed: 21524568]
- Hubbard A, Jamshidian F, Jewell N. Adjusting for perception and unmasking effects in longitudinal clinical trials. *International Journal of Biostatistics*. 2012; 8(2):Article 7.
- 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]
- Jamshidian F, Hubbard AE, Jewell NP. Accounting for perception, placebo and unmasking effects in estimating treatment effects in randomized clinical trials. *Statistical Methods in Medical Research*. 2011; 23:293–307. [PubMed: 21908418]

- Molenberghs, G.; Kenward, MG. Missing Data in Clinical Trials. New York: Wiley; 2007.
- 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]
- Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period —application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986; 7:1393–1512.
- Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*. 2000:6–10.
- 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*. 1983; 70:41–55.
- Roy J. Randomized treatment-belief trials. *Contemporary Clinical Trials*. 2012; 33:172–177. [PubMed: 21989161]
- 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.
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*. 1999; 94:1096–1146.
- Stefanski LA, Boos DD. The calculus of M-estimation. *The American Statistician*. 2002; 56:29–38.
- 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.; Gruber, S. Technical Report. UC Berkeley Division of Biostatistics; 2011. Targeted minimum loss based estimation of an intervention specific mean outcome. <http://www.bepress.com/ucbbiostat/paper290>
- van der Laan, MJ.; Robins, JM. Unified Methods for Censored Longitudinal Data and Causality. New York: Springer-Verlag; 2003.
- van der Laan, MJ.; Rose, S. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer; New York: 2011.
- van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; Cambridge: 1998.
- Zhang Z, Kotz RM, Wang C, Ruan S, Ho M. A causal model for joint evaluation of placebo and treatment-specific effects in clinical trials. *Biometrics*. 2013; 69:318–327. [PubMed: 23432119]
- Zhang Z, Nie L, Soon G, Zhang B. Sensitivity analysis in non-inferiority trials with residual inconstancy after covariate adjustment. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*. 2014; 63:515–538.

Table 1

Subject availability at 6 and 12 months in the weight loss study.

Available		Group		
6 months	12 months	Active	Placebo	Combined
yes	yes	198	199	397
yes	no	12	19	31
no	no	16	20	36
no	yes	9	6	15
Column Total		235	244	479

Table 2

Summary of the available %EWL data (excluding missing data) in the weight loss study.

Time	Statistic	Group	
		Active	Placebo
6 months	<i>n</i>	210	218
	mean	24.7	17.4
	sd	16.9	18.0
12 months	<i>n</i>	207	205
	mean	36.3	22.7
	sd	28.5	24.6

Table 3

Summary of the available treatmentality data (excluding missing data) in the weight loss study. Each numerical entry is a percentage within a treatment group.

Time	Answer	Group	
		Active	Placebo
1 week	active	54.4	54.6
	placebo	31.0	37.0
	unknown	14.6	8.4
6 months	active	73.3	37.2
	placebo	11.4	51.8
	unknown	15.2	11.0
12 months	active	76.5	21.8
	placebo	6.5	62.4
	unknown	17.0	15.8

Table 4

Main analysis of the weight loss study data: point estimates and standard errors for $\mu(t, \bar{s})$ ($t = 0, 1$; $\bar{s} = \bar{0}, \bar{1}$) and the associated causal effects (δ 's) and interaction (τ), obtained using the naive method and the GC/SR, IPW, AIPW and TMLE methods under the BIC-based models (see Section 4 for details).

Parameter	Point Estimate						Standard Error					
	naive	GC/SR	IPW	AIPW	TMLE1	TMLE2	naive	GC/SR	IPW	AIPW	TMLE1	TMLE2
$\mu(1, \bar{1})$	47.2	39.7	43.2	40.4	40.6	40.9	3.0	2.8	3.0	2.8	2.9	2.8
$\mu(0, \bar{1})$	34.4	19.5	21.1	18.9	17.6	19.6	3.4	2.9	4.1	3.7	6.0	4.2
$\mu(1, \bar{0})$	16.2	25.3	20.9	24.0	23.2	22.5	5.6	3.6	10.1	3.8	6.5	5.6
$\mu(0, \bar{0})$	14.5	16.1	16.6	16.4	16.6	16.9	2.3	2.0	2.6	2.0	2.1	2.0
$\delta(, \bar{1})$	12.8	20.2	22.1	21.5	23.0	21.3	4.6	4.1	5.3	4.7	6.6	5.0
$\delta(, \bar{0})$	1.6	9.1	4.3	7.6	6.6	5.5	6.0	4.1	11.0	4.2	6.8	5.9
$\delta(1, \cdot)$	31.0	14.4	22.3	16.4	17.4	18.4	6.6	4.6	10.6	4.7	7.2	6.3
$\delta(0, \cdot)$	19.8	3.4	4.5	2.5	1.0	2.7	4.1	3.4	4.9	4.2	6.3	4.6
τ	11.2	11.1	17.8	13.9	16.4	15.7	7.8	5.8	12.3	6.4	9.5	7.8

Table 5

Simulation results for different blinding mechanisms: empirical bias and standard deviation for estimating $\mu(t, \bar{s})$ ($t = 0, 1; \bar{s} = \bar{0}, \bar{1}$) and the associated causal effects (δ 's) and interaction (τ), in various scenarios of treatmentality distribution (see Section 5 for details). Each entry is based on 10000 replicates.

Parameter	Bias					Standard Deviation						
	naive	GC/SR	IPW	AIPW	TMLE1	TMLE2	naive	GC/SR	IPW	AIPW	TMLE1	TMLE2
estimated BIC models												
$\mu(1, \bar{1})$	4.2	0.0	0.1	0.0	0.0	0.1	3.0	2.7	3.6	2.9	2.9	2.9
$\mu(0, \bar{1})$	14.4	0.0	0.4	0.0	0.0	1.2	3.4	2.9	8.8	3.8	5.4	4.3
$\mu(1, \bar{0})$	-9.7	0.0	-0.7	0.0	0.0	-0.1	4.8	3.2	15.1	4.8	6.7	6.1
$\mu(0, \bar{0})$	-1.9	0.0	0.0	0.0	0.0	0.0	2.7	2.2	4.2	2.6	2.7	2.7
$\delta(\cdot, \bar{1})$	-10.2	0.0	-0.4	0.0	0.1	-1.1	4.5	3.9	9.7	4.8	6.2	5.2
$\delta(\cdot, \bar{0})$	-7.8	0.0	-0.7	0.0	0.0	-0.1	5.6	3.8	16.2	5.5	7.2	6.6
$\delta(1, \cdot)$	14.0	0.0	0.7	0.0	0.0	0.2	5.8	4.2	15.5	5.6	7.3	6.7
$\delta(0, \cdot)$	16.3	0.0	0.4	0.0	-0.1	1.2	4.4	3.7	9.8	4.7	6.0	5.0
τ	-2.4	0.0	0.3	0.0	0.1	-1.0	7.3	5.5	19.0	7.2	9.4	8.3
perfect blinding												
$\mu(1, \bar{1})$	3.8	0.0	0.1	0.0	0.0	0.1	3.0	2.7	3.4	2.8	2.9	2.8
$\mu(0, \bar{1})$	14.5	0.0	0.4	0.0	0.0	1.2	3.4	2.9	8.3	3.8	5.4	4.3
$\mu(1, \bar{0})$	-9.0	0.0	-0.6	0.0	0.0	-0.1	4.5	3.1	13.6	4.5	6.2	5.7
$\mu(0, \bar{0})$	-1.9	0.0	0.0	0.0	0.0	0.0	2.9	2.3	4.2	2.7	2.8	2.8
$\delta(\cdot, \bar{1})$	-10.7	0.0	-0.4	0.0	0.0	-1.1	4.6	4.0	9.1	4.8	6.2	5.2
$\delta(\cdot, \bar{0})$	-7.1	0.0	-0.6	-0.1	0.0	-0.2	5.3	3.8	14.8	5.3	6.8	6.3
$\delta(1, \cdot)$	12.8	0.0	0.6	0.0	-0.1	0.2	5.5	4.1	14.0	5.4	6.9	6.4
$\delta(0, \cdot)$	16.4	0.0	0.4	0.0	0.0	1.1	4.5	3.7	9.3	4.7	6.1	5.1
τ	-3.6	0.0	0.2	0.0	0.0	-1.0	7.1	5.5	17.4	7.1	9.2	8.1
severe unblinding												
$\mu(1, \bar{1})$	4.6	0.0	0.1	0.0	0.0	0.2	3.0	2.7	4.7	3.6	3.6	3.4
$\mu(0, \bar{1})$	14.4	0.0	0.4	0.0	0.0	1.2	3.4	2.9	8.4	3.9	5.4	4.3

Parameter	Bias					Standard Deviation						
	naive	GC/SR	IPW	AIPW	TMLE1	TMLE2	naive	GC/SR	IPW	AIPW	TMLE1	TMLE2
$\mu(1, \bar{0})$	-10.2	0.0	-0.8	0.0	0.0	-0.2	5.0	3.3	18.0	6.2	9.8	8.2
$\mu(0, \bar{0})$	-1.9	0.0	-0.1	0.0	0.0	0.0	2.7	2.1	3.7	2.5	2.5	2.5
$\mathcal{E}(\cdot, \bar{1})$	-9.8	0.0	-0.4	0.0	0.0	-1.1	4.6	4.0	9.8	5.3	6.5	5.5
$\mathcal{E}(\cdot, \bar{0})$	-8.3	0.0	-0.8	0.0	0.1	-0.2	5.7	3.9	18.8	6.7	10.1	8.6
$\mathcal{E}(1, \cdot)$	14.8	0.1	0.9	0.0	0.0	0.4	5.8	4.3	18.6	7.1	10.4	8.9
$\mathcal{E}(0, \cdot)$	16.3	0.1	0.5	0.0	0.1	1.3	4.3	3.6	9.2	4.6	5.9	5.0
τ	-1.5	0.0	0.4	0.0	0.0	-0.9	7.3	5.6	21.2	8.5	12.0	10.2
more unknowns												
$\mu(1, \bar{1})$	4.4	0.0	0.2	0.0	0.1	0.2	3.9	3.1	4.7	3.4	3.6	3.5
$\mu(0, \bar{1})$	14.5	0.0	0.8	0.0	-0.1	1.6	4.2	3.2	8.5	4.4	7.4	5.5
$\mu(1, \bar{0})$	-9.7	0.0	-1.9	0.0	0.0	0.0	7.0	4.0	18.0	6.1	11.1	9.5
$\mu(0, \bar{0})$	-1.9	0.0	0.0	0.0	0.0	0.0	3.3	2.5	5.0	3.1	3.2	3.2
$\mathcal{E}(\cdot, \bar{1})$	-10.1	0.0	-0.6	0.1	0.1	-1.4	5.8	4.4	9.8	5.5	8.2	6.5
$\mathcal{E}(\cdot, \bar{0})$	-7.8	0.0	-1.9	0.0	0.0	0.0	7.7	4.6	19.5	6.9	11.6	10.0
$\mathcal{E}(1, \cdot)$	14.1	0.0	2.1	0.1	0.1	0.2	8.0	5.0	18.6	7.0	11.7	10.1
$\mathcal{E}(0, \cdot)$	16.4	0.0	0.7	0.0	-0.1	1.6	5.4	4.0	9.9	5.3	8.1	6.3
τ	-2.3	0.0	1.3	0.1	0.2	-1.4	9.6	6.3	21.7	8.8	14.2	11.9
more optimism												
$\mu(1, \bar{1})$	4.1	0.0	0.1	0.0	0.0	0.1	2.7	2.4	3.2	2.6	2.6	2.5
$\mu(0, \bar{1})$	14.6	0.0	0.4	0.0	0.0	1.0	3.0	2.6	7.7	3.4	4.5	3.6
$\mu(1, \bar{0})$	-9.9	0.0	-1.9	0.0	0.1	0.1	6.5	4.0	17.8	6.4	12.0	10.0
$\mu(0, \bar{0})$	-1.9	0.0	-0.1	0.0	0.0	0.0	3.7	2.8	5.5	3.6	3.9	3.8
$\mathcal{E}(\cdot, \bar{1})$	-10.4	0.0	-0.4	0.0	0.0	-0.9	4.1	3.5	8.4	4.3	5.2	4.4
$\mathcal{E}(\cdot, \bar{0})$	-7.9	0.0	-1.8	0.0	0.1	0.1	7.5	4.8	19.3	7.3	12.5	10.6
$\mathcal{E}(1, \cdot)$	14.0	0.0	2.0	0.0	-0.1	0.0	7.1	4.7	18.1	6.9	12.3	10.3
$\mathcal{E}(0, \cdot)$	16.5	0.0	0.5	0.0	0.0	1.0	4.8	3.8	9.5	5.0	5.9	5.2
τ	-2.5	0.0	1.4	0.0	-0.1	-1.0	8.6	6.0	20.9	8.5	13.5	11.4

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Parameter	Bias						Standard Deviation					
	naïve	GC/SR	IPW	AIPW	TMLE1	TMLE2	naïve	GC/SR	IPW	AIPW	TMLE1	TMLE2
$\mu(1, \bar{1})$	4.4	0.1	0.2	0.1	0.1	0.2	3.6	3.1	4.5	3.5	3.6	3.5
$\mu(0, \bar{1})$	14.3	0.0	0.6	0.0	0.0	1.5	4.2	3.4	9.1	4.6	7.3	5.6
$\mu(0, 1)$	-9.5	0.0	-0.4	0.0	0.0	-0.1	3.9	2.8	12.4	3.9	4.6	4.4
$\mu(0, 0)$	-1.9	0.0	0.0	0.0	0.0	0.0	2.3	1.9	3.1	2.1	2.1	2.1
$\delta^{\ast}, \bar{1}$	-10.0	0.1	-0.4	0.1	0.1	-1.3	5.5	4.6	10.3	5.8	8.1	6.6
$\delta^{\ast}, 0$	-7.6	0.0	-0.4	0.0	0.0	-0.1	4.5	3.3	13.2	4.4	5.1	4.8
$\delta^{\ast}, 1$	13.9	0.1	0.5	0.0	0.1	0.3	5.3	4.2	13.2	5.3	5.9	5.6
$\delta^{\ast}, 0$	16.3	0.0	0.6	0.0	0.0	1.6	4.8	3.9	9.6	5.1	7.6	6.0
τ	-2.3	0.1	0.0	0.1	0.1	-1.3	7.2	5.7	16.9	7.3	9.6	8.1

less optimism