Statistics
in Medicine

Research Article

Received 13 April 2015, Accepted 28 September 2015 Published online 19 October 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6768

Estimating causal effects for multivalued treatments: a comparison of approaches

Ariel Linden,[a,c*†] S. Derya Uysal,[b] Andrew Ryan[c] and John L. Adams[d]

Interventions with multivalued treatments are common in medical and health research, such as when comparing the efficacy of competing drugs or interventions, or comparing between various doses of a particular drug. In recent years, there has been a growing interest in the development of multivalued treatment effect estimators using observational data. In this paper, we compare the performance of commonly used regression-based methods that estimate multivalued treatment effects based on the unconfoundedness assumption. These estimation methods fall into three general categories: (i) estimators based on a model for the outcome variable using conventional regression adjustment; (ii) weighted estimators based on a model for the treatment assignment; and (iii) 'doubly-robust' estimators that model both the treatment assignment and outcome variable within the same framework. We assess the performance of these models using Monte Carlo simulation and demonstrate their application with empirical data. Our results show that (i) when models estimating both the treatment and outcome are correctly specified, all adjustment methods provide similar unbiased estimates; (ii) when the outcome model is misspecified, regression adjustment performs poorly, while all the weighting methods provide unbiased estimates; (iii) when the treatment model is misspecified, methods based solely on modeling the treatment perform poorly, while regression adjustment and the doubly robust models provide unbiased estimates; and (iv) when both the treatment and outcome models are misspecified, all methods perform poorly. Given that researchers will rarely know which of the two models is misspecified, our results support the use of doubly robust estimation. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: multivalued treatments; regression adjustment; propensity score weighting; doubly robust; inverse probability weights; observational studies

1. Introduction

Interventions with multivalued treatments are common in medical and health research. Multivalued treatments may include more than two discrete conditions (e.g., comparing the efficacy of competing drugs or interventions) or multiple levels of one treatment (e.g., various doses of a particular drug). In experimental studies, outcomes for multivalued treatments may be analyzed by simply regressing the outcome on a set of indicator variables representing each treatment, followed by contrasts between the treatment variables to estimate treatment effects. This minimal setup is sufficient to provide unbiased treatment effect estimates when subjects are randomized. However, when analyzing observational data, investigators estimate treatment effects by applying causal-inferential methods to control for confounders.

In recent years, there has been a growing interest in the development of multivalued treatment effect estimators using observational data. The seminal work of Imbens [1] and Lechner [2] gave rise to this burgeoning area by extending Rosenbaum and Rubin's (1983) propensity score framework for binary treatments to multivalued treatments. Subsequently, several methods designed for binary treatments–including regression, matching, weighting, subclassification on covariates and stratification–have been

[a] *Linden Consulting Group, LLC, Ann Arbor, MI, U.S.A.*
[b] *Department of Economics and Finance, IHS, Vienna, Austria*
[c] *Department of Health Management & Policy, University of Michigan School of Public Health, Ann Arbor, MI, U.S.A.*
[d] *Kaiser Permanente, Center for Effectiveness and Safety Research, Pasadena, CA, U.S.A.*
**Correspondence to: Ariel Linden, DrPH Linden Consulting Group, LLC 1301 North Bay Drive Ann Arbor, MI U.S.A. 48103.*
[†] *E-mail: alinden@lindenconsulting.org*

reformulated to accommodate multivalued treatments [3–10]. These estimators have valuable applications in new priority areas of health services research, particularly comparative effectiveness research, which seeks to compare the effects of multiple therapies [11].

This paper contributes to the literature by comparing the performance of commonly used regression-based methods that estimate multivalued treatment effects based on the unconfoundedness assumption in pretest–posttest studies. These estimation methods fall into three general categories: (i) estimators based on a model for the outcome variable using conventional regression adjustment (RA); (ii) estimators based on a model for the treatment assignment, using inverse probability of treatment weighting (IPTW) [12, 13] and marginal mean weighting through stratification (MMWS) [9, 14]; and (iii) 'doubly-robust' estimators that model both the treatment assignment and outcome variable within the same framework, using an augmented IPTW approach (A-IPTW) [15–17] and IPTW combined with RA (IPTW-RA) [8, 18, 19]. We examine these models using both Monte Carlo simulation and empirical data from a disease management program for patients with congestive heart failure that were exposed to one of three study arms. From the literature on asymptotics of those estimators, if the treatment and outcome models are correctly specified, regression adjustment is more efficient than A-IPTW and IPTW-RA; and A-IPTW and IPTW-RA are more efficient estimators than IPTW [18, 20]. To the best of our knowledge, there is no general result on the relative efficiency of A-IPTW versus IPTW-RA. Similarly, MMWS has only been contrasted with IPTW [9]. Therefore, the main goals of the Monte Carlo study are the following: (i) to investigate and to contrast the finite sample properties of several estimation methods for correctly specified models; (ii) to evaluate the finite sample properties of these approaches under model misspecification; and (iii) to investigate the doubly robustness property of A-IPTW and IPTW-RA. The empirical example demonstrates the application of these estimators using data from a disease management program evaluation.

While our paper covers a wide variety of methods existing in the literature, we do not include propensity score matching and subclassification on covariates. In the binary treatment case, matching on the propensity score serves a similar purpose to that of weighting and stratification. However, matching has unique challenges when extended to the multiple treatment case. Conceptually, multiple treatment matching either attempts binary treatment matching for all pairwise comparisons or it searches for triplets (or multuplets) that match across the multiple treatment arms. The all-pairwise case is complicated by the potentially different supports of the pairwise comparisons. The matched multuplets case suffers from the 'curse of dimensionality' that can make finding enough matched sets difficult unless the available dataset is very large. Although propensity scoring is a great aid in both these cases, the likely greater effects of lack of common support in matching suggest that a proper evaluation of multiple treatment matching would require a different simulation study design than used here. The second approach, which is not covered in our paper, is subclassification on covariates, as described by Cattaneo and Farrell [10]. As the method is intrinsically nonparametric and does not entail estimating the propensity score or outcome model, it would also require a different simulation study. Cattaneo and Farrell [10], however, provide a comprehensive Monte Carlo comparison study of stratification on covariates versus matching methods for binary treatments.

This paper is organized as follows: Section 2 describes the potential outcomes framework applied to multivalued treatments. Section 3 introduces the various methods that are compared in the paper. Section 4 details the construction and results of the Monte Carlo simulation. Section 5 describes the empirical study and reports the results, and Section 6 provides discussion and conclusions.

2. Potential outcomes framework for multivalued treatments

Consider N units that are drawn from a large population. For each individual i, $i = 1, \dots, N$ in the sample, the triple (Y_i, T_i, X_i) is observed. Y_i is the outcome variable, T_i is the multivalued treatment variable, which takes the integer values between 0 and K , and X_i represents the vector of pre-treatment covariates. $D_{it}(T_i)$ is the indicator of receiving the treatment t for individual i:

$$D_{it}(T_i) = \begin{cases} 1, & \text{if} \quad T_i = t \\ 0, & \text{otherwise.} \end{cases}$$

For each individual, there is a set of potential outcomes $(Y_{i0}, \dots, Y_{iK})$. Y_{it} denotes the potential outcome for each individual i, for which $T_i = t$ where $t \in \mathfrak{T} = \{0, \dots, K\}$. Only one of the potential outcomes is observed, depending on the treatment status.[‡] Adopting the potential outcomes framework of Rubin [21], the observed outcome, Y_i, can be written in terms of treatment indicator, $D_{it}(T_i)$, and the potential outcomes, Y_{it}:

$$Y_i = \sum_{t=0}^{K} D_{it}(T_i) Y_{it}. \tag{1}$$

Thus, the individual-level treatment effect of treatment level m versus l is $Y_{im} - Y_{il}$; the difference of these two potential outcomes. The population average treatment effect is given by the difference in the means of the two potential outcomes:

$$\Delta_{ml} = \mathrm{E}\left[Y_{im} - Y_{il}\right] = \mu_m - \mu_l. \tag{2}$$

In a randomized experiment, Δ_{ml} can be estimated using the sample means of observed outcomes. In an observational study, however, estimation of Δ_{ml} requires additional conditioning on X_i, which is assumed to contain all confounders associated with both the treatment assignment mechanism and potential outcomes. By conditioning on X_i, we assume that treatment assignment is as good as randomly assigned–thereby replicating the randomization process. This assumption, also called weak unconfoundedness, as defined by Imbens [1], can be formally stated as follows:

$$Y_{it} \perp D_{it}(T_i) | X_i, \forall t \in \mathfrak{T},$$

where $\perp$ denotes orthogonality or independence. This assumption requires that all determinants of treatment level and of the outcome variable are observed. Clearly, this is a strong assumption and generally requires a rich dataset, in application. A second assumption that is typically considered in conjunction with unconfoundedness is that of a complete overlap in the distribution of covariates between treatment groups. More specifically as follows:

$$0 < \Pr\left[T_i = t \,|X_i = x\right], \forall t \in \mathfrak{T} \text{ and } \forall x \text{ in the support of } X.$$

Rosenbaum and Rubin [22] refer to the combination of unconfoundedness and overlap as strong ignorability. Hence, under these assumptions, the conditional expectation of potential outcome for treatment level t identified by conditional expectation of observed outcomes of individuals receiving treatment t [23]:

$$\begin{aligned}\mathrm{E}\left[Y_{it}\middle| X_i\right] &= \mathrm{E}\left[Y_{it}\middle| D_{it}(T_i), X_i\right] = \mathrm{E}\left[Y_i\middle| D_{it}(T_i), X_i\right] \\ &= \mathrm{E}\left[Y_i\middle| T_i, X_i\right].\end{aligned} \tag{3}$$

Thus, the unconditional means can be estimated by averaging these conditional means, that is, $\mu_t \equiv \mathrm{E}\left[Y_{it}\right] = \mathrm{E}\left[\mathrm{E}\left[Y_{it}\middle| X_i\right]\right]$. For high dimensional X_i, Imbens [1] introduced the generalized propensity score (GPS) to serve as a practical alternative to conditioning directly on X_i in the case of multivalued treatments. The GPS is defined as the conditional probability of receiving a particular level of the treatment given the pretreatment variables, such that:

$$r(t, x) \equiv \Pr\left[T_i = t \,|X_i = x\right] = \mathrm{E}\left[D_{it}(T_i)\middle| X_i = x\right]. \tag{4}$$

Identification of potential outcomes' means is also possible, as in the binary treatment case, by weighting observed outcomes by the conditional probability of the received treatment [1]:

$$\mathrm{E}\left[\frac{Y_i D_{it}(T_i)}{r(t, X_i)}\right] = \mathrm{E}\left[Y_{it}\right]. \tag{5}$$

In practice GPS, $r(t, X_i)$ is usually not known but can be estimated by discrete response models if the multivalued treatment does not have a logical ordering, or by ordered response models if the treatment corresponds to ordered levels [1].

[‡] *Note that if K contains only two values, treatment reverts to the binary case.*

3. Approaches for causal inference in multivalued treatments

In this section, we briefly describe the approaches to causal inference in multivalued treatments, which are valid if the uncounfoundedness and overlap assumptions are satisfied, with an emphasis on the implementation of the described estimators.[§] We limit our discussion to the methods, which are compared in this paper. It draws heavily on the technical discussion in the references provided, and we suggest that the reader use those references for more background on, and formal derivations of, some of the properties of the estimators described here.

3.1. Regression adjustment

If the uncounfoundedness assumption stated in the previous section is satisfied, that is, if we observe all the factors associated with the treatment status as well as outcome variable, the conditional mean of each potential outcome can be identified by the conditional mean function of the observed outcome using the units that were exposed to the relevant treatment level (Equation (3)). Thus, multiple regression is an estimator based on a model for the outcome variable. In this approach, the outcome is regressed on a set of covariates separately for each treatment level, after which the predicted outcomes for each subject and treatment level are computed using data only from the individuals receiving the relevant treatment level. The average of these predicted values estimates the potential outcome means, which can then be contrasted to estimate average treatment effects. More formally, using regression adjustment and assuming unconfoundedness, we can define the conditional mean functions of the potential outcomes as follows:

$$m_t(X_i) = \mathrm{E}\left[Y_{it}\middle|X_i\right] = \mathrm{E}\left[Y_i\middle|T_i = t, X_i\right] = \beta_{0t} + X_i'\beta_{1t}, \quad \forall t \in \mathfrak{T} \tag{6}$$

Average treatment effects can then be estimated by contrasting estimated potential outcome means between any two treatment levels:

$$\begin{aligned}\hat{\Delta}^{RA}_{ml} &= \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\beta}_{0m} + X_i'\hat{\beta}_{1m}\right) - \left(\hat{\beta}_{0l} + X_i'\hat{\beta}_{1l}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\hat{m}_m(X_i) - \hat{m}_l(X_i)\right)\end{aligned}, \tag{7}$$

where m and l may represent any two treatment levels in the set. While RA is a widely-used technique, possible misspecifications of the functional form of the outcome model could bias the treatment effect estimate [24]. Additionally, regression relies on extrapolation for estimation when the distribution of covariates between treatment groups is substantially different. This last issue has provided support for the use of weighting techniques as an alternative evaluation approach to RA, as they allow the investigator to directly assess covariate balance between treatment groups.

3.2. Inverse probability of treatment weighting

The concept of inverse probability weighting originated in survey research over 60 years ago to adjust for imbalances in sampling pools [25] and continues to be regularly used in complex survey designs. Over the years, this weighting concept has also been extended to the study of treatment effects in observational studies [see for example, 12, 13], where the weighting estimators are used to model the IPTW. Using the sample counterpart of Equation (5), the estimator for the average treatment effect is given by the following:

$$\hat{\Delta}^{IPTW1}_{ml} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_iD_{im}(T_i)}{\hat{r}(m, X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{Y_iD_{il}(T_i)}{\hat{r}(l, X_i)} = \hat{\mu}_m - \hat{\mu}_l, \tag{8}$$

where $\hat{r}(t, X_i)$ is the estimated GPS, and m and l may represent any two treatment levels in the set [1].

[§] *We do not discuss the variance estimation or asymptotic properties of the methods; however, interested readers can refer to* Cattaneo [17] *for a detailed and formal discussion on efficiency of the methods.*

As in the binary treatment case, one can normalize the weights such that they add up to one, in order to improve the finite sample properties [19]. The treatment effect estimator based on these normalized weights can be written as follows:

$$\hat{\tau}_{ml}^{IPTW2} = \left[\sum_{i=1}^{N} \frac{Y_i D_{im}(T_i)}{\hat{r}(m, X_i)} \bigg/ \sum_{i=1}^{N} \frac{D_{im}(T_i)}{\hat{r}(m, X_i)} \right] - \left[\sum_{i=1}^{N} \frac{Y_i D_{il}(T_i)}{\hat{r}(l, X_i)} \bigg/ \sum_{i=1}^{N} \frac{D_{il}(T_i)}{\hat{r}(l, X_i)} \right] \tag{9}$$

As Cattaneo [17] has shown, $\hat{\tau}_{ml}^{IPTW2}$ actually emerges from the generalized method of moments (GMM) representation of the treatment effects. An advantage that IPTW estimators hold over RA is that the degree of overlap in the distribution of covariates between treatment levels (i.e., covariate balance) can be directly assessed using numeric summaries (such as standardized differences or variance ratios) and graphical displays (such as box plots or Q–Q plots), as observed covariate balance is an essential criterion for helping to ensure that treatment effects are valid in studies of treatment effects [26, 27]. However, a limitation of the IPTW framework is that treatment effect estimates can be distorted when the overlap assumption is violated, or more specifically, when individuals with no counterfactual information under an alternative treatment condition receive a nonzero weight [14]. Another limitation of IPTW is that it can perform poorly when the weights for a few subjects are very large. In this situation, the treatment effect estimates may become very imprecise because of large standard errors [28].

3.3. Marginal mean weighting through stratification

Recently, an approach called marginal mean weighting through stratification [9, 14] has been introduced that combines elements of both propensity score stratification and IPTW. In general, this first entails stratifying the analytic sample into quantiles of the generalized propensity score, and then generating a weight for each individual based on their corresponding stratum and treatment assignment. The stratification reduces bias in the observed covariates used to create the propensity score [29], and the weighting standardizes each treatment group to the target population [30].

In the multivalued treatment setting, the MMWS approach is conducted as follows: first, the GPS is estimated either by an ordered or mutinomial response models. Next, each GPS is stratified into equal sized quantile categories. If an ordered response model is used, stratification is based on the estimated probability of the base category, and if a multinomial response model is used, the sample is stratified separately for each of the estimated probabilities. Typically, investigators divide the data into five strata, as it has been shown that stratifying the propensity score into quintiles can remove over 90% of the selection bias [22, 29]. Moreover, in large samples, further bias reduction may be achieved by adding additional strata. Next, the marginal mean weights are computed based on the formula by Hong [14]:

$$\text{MMW} = \frac{n_{s_t} \times \widehat{\Pr}\,[T = t]}{n_{T=t,s_t}} \tag{10}$$

where $\widehat{\Pr}\,[T = t]$ is the estimated probability of assignment to treatment group t, that is, the proportion of those actually receiving treatment t in the population, n_{s_z} is the number of units in stratum s_t constructed on the estimated probability of treatment level t, and $n_{T=t,s_t}$ is the number of units in stratum s_t who were actually assigned to treatment t.[¶] Thus, the weight is proportional to the ratio of the number of individuals in a given strata to the number of individuals within that strata actually receiving the treatment. The unconditional mean is estimated in the usual fashion, with the MMWS weights used as sampling weights:

$$\hat{\mu}_t^{MMW} = \frac{\sum_{i=1}^{N} \text{MMW}_i Y_i D_{it}(T_i)}{\sum_{i=1}^{N} \text{MMW}_i D_{it}(T_i)}. \tag{11}$$

Average treatment effects can then be estimated by contrasting estimated potential outcome means between any two treatment levels.

Marginal mean weighting through stratification has been shown to be more accurate than IPTW in estimating outcomes in the binary treatment case. Huang [31] used both techniques and found that the

[¶] *Note that in an ordered treatment case there is no subscript for s, as the sample is stratified only once on the probability of the base category.*

IPTW results were much more variable, and in many cases, did not agree with the other two methods applied to the data (the stratification approach and hierarchical outcome regression). Similarly, Hong [9] found through a comprehensive set of simulations that MMWS achieved lower bias and mean squared error than IPTW when the propensity score model was misspecified. Hong [9] attributes the better performance of MMWS over IPTW to the stratification component of the procedure. She argues that even when the propensity score is misspecified, membership in the propensity score stratum remains consistent for individuals in their respective treatment groups. Because MMWS is estimated as a ratio of the sample sizes within each stratum, the MMWS estimate of treatment effect will therefore remain robust.

3.4. Augmented Inverse probability of treatment weighting

A class of estimators has evolved to model both the probability of treatment and the outcome simultaneously within the same framework, providing asymptotically unbiased estimates when only one of the two models is correctly specified. These estimators are called ‘doubly robust’ because they provide the investigator two opportunities to derive consistent treatment effects [13, 16]. While there are several different doubly robust methods available [see 32, for a review of various DR approaches], the most commonly-used approach is that credited to Robins and colleagues [13, 15, 16]. This estimator incorporates an augmentation term in the IPTW estimator so that the treatment effect estimator stays consistent even if the GPS model is misspecified [15, 20]. If the GPS model is correctly specified, the augmentation term goes to zero in large samples [33]. The unconditional mean is thus estimated as follows:

$$\hat{\mu}_t^{A-IPTW} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Y_i D_{it}(T_i)}{\hat{r}(t,X_i)} - \frac{D_{it}(T_i) - \hat{r}(t,X_i)}{\hat{r}(t,X_i)}\hat{m}_t(X_i)\right]. \tag{12}$$

Essentially, $\hat{\mu}_t^{A-IPTW}$ corresponds to Cattaneo’s efficient influence function estimator [17]. The model can be operationalized in a three-step process: first, the parameters of the generalized propensity score model are estimated and the IPT weights computed. Next, separate regression models of the outcome are estimated for each treatment level, and the treatment-specific predicted outcomes for each individual are obtained. Finally, unconditional means are estimated as in Equation (12), using the estimated GPS from the first step, $\hat{r}(t, X_i)$, as well as the estimated conditional mean functions, $\hat{m}_t(X_i)$. The contrasts of these weighted averages provide the estimates for the ATE. The estimation procedure can also be reduced to one-step estimation if the GMM approach is used. The GMM framework makes it easier to derive standard errors, which are adjusted for estimation errors originating from the estimated GPS and outcome model.

Simulation evidence for the binary counterpart of this estimator suggests that, while the A-IPTW model may be less efficient than regression adjustment when the outcome function is correctly specified, the A-IPTW is more robust against misspecification compared with the single-model methods due to the doubly robustness property [16].

3.5. Inverse probability of treatment weighted regression adjustment

The weighted regression estimator simultaneously estimates the GPS and the outcome model while using weights that are equal to the inverse of the GPS. For certain classes of models used to estimate the outcome, this method has been shown to be doubly robust [18, 34]. Like the A-IPTW, the IPTW-RA model can also be operationalized in a three-step process: First, the parameters of the generalized propensity score model are estimated, and the IPT weights are computed as $\frac{D_{it}(T_i)}{\hat{r}(t,X_i)}$ for each level of treatment. Next, using the estimated IPTW, the outcome models in Equation (6) are fitted by a weighted regression for each treatment level, and treatment-specific predicted outcomes for each individual are obtained using the estimated coefficients from this weighted regression. Finally, the means of the treatment-specific predicted outcomes are computed. The contrasts between these averages provide the estimates of the ATEs [see 19, for an application of this method]. One can also rewrite the estimation procedure as a one step estimation within a GMM framework. As in the previous method, the major advantage of the GMM approach is in deriving standard errors, which automatically account for the estimation error from the estimated GPS [8, 19].

An important consideration for all methods that assume unconfoundedness is the choice of covariates selected for inclusion in the modeling process (in both the GPS and outcome model). There is a wide variety of methods available for selecting covariates, ranging from *ad-hoc* manual selection to fully automated data-driven techniques. Perhaps the most intuitive program in this class is that introduced by

Catteneo *et al.* [35], which automatically processes the iterative tasks that an investigator would otherwise perform manually. The program `bfit` written for use in Stata, generates a series of candidate models ranging from a model including a single covariate to a model that includes a fully interacted polynomial of the order specified by the user. The best-fitting model is then determined by the BIC or the AIC. For those investigators interested in a more sophisticated approach to covariate selection, Farrell [36] introduces a method based on the group lasso, which is particularly well-suited to multivalued treatments with sparse data.

4. Monte Carlo simulations

4.1. *Basic simulation design*

We model four scenarios, where: (i) both the treatment (i.e., the GPS) and outcome models are correctly specified; (ii) only the treatment model is correctly specified; (iii) only the outcome model is correctly specified; and (iv) neither model is correctly specified. In each scenario, we draw 10,000 replications from the data-generating process, repeated for sample sizes of 500 and 2000. In each replication, we perform estimation and inference for the means of three treatment levels ($t \in 0, 1, 2$). At each iteration, for each model and treatment level, we record the point estimates, the standard error, the squared error, and a binary indicator noting whether the null hypothesis is rejected that the parameter equals its true value, based on the model's point estimate and standard error.

4.2. *Data generating process*

Following Cattaneo *et al.* [35], we draw samples from four data generating processes (DGPs). In all four data generating processes, the GPSs are generated from a multinomial logit, and the outcome variable Y is drawn from a Weibull distribution conditional on the treatment level t and the two covariates X_1 and X_2. Both covariates are drawn from a uniform distribution over $(-0.5, 0.5)$.

4.2.1. Data for the treatment model. As in Cattaneo *et al.* [35], there are three treatment levels ($t \in \{0, 1, 2\}$), and the true propensity score is a multinomial logit (with treatment level 0 as base level),

$$\Pr\left[T_i = 0 \left| X_i\right.\right] = \frac{1}{q_i}, \quad \Pr\left[T_i = 1 \left| X_i\right.\right] = \frac{ex_{1i}}{q_i}, \quad \Pr\left[T_i = 2 \left| X_i\right.\right] = \frac{ex_{2i}}{q_i},$$

where

$$ex_{1i} = \exp\left\{1.5\left(-.2 + X_{1i} + X_{2i}\right)\right\},$$
$$ex_{2i} = \exp\left\{1.2\left(-.1 + X_{1i} + X_{2i}\right)\right\},$$

and $q_i = 1 + ex_{1i} + ex_{2i}$. Given the probabilities and the [0, 1] uniform random variable u_i,

$$T_i = \begin{cases} 0, & \text{if} \quad u_i \leqslant \Pr\left[T_i = 0 \left| X_i\right.\right] \\ 1, & \text{if} \quad \Pr\left[T_i = 0 \left| X_i\right.\right] < u_i \leqslant \Pr\left[T_i = 0 \left| X_i\right.\right] + \Pr\left[T_i = 1 \left| X_i\right.\right] \\ 2, & \text{otherwise} \end{cases} .$$

When using a standard multinomial logit model, the GPS is modeled as a function of X_1 and X_2 in the correctly specified case and is modeled as a function of only X_1 in the misspecified case (i.e., X_2 is omitted).

4.2.2. Data for the outcomes model. A Weibull distribution for Y conditional on X was chosen because it is asymmetric, continuous, and specifies the mean as a nonlinear function of the parameters of the distribution. The Weibull distribution had a scale parameter η and a shape parameter θ, with a mean $\eta\Gamma\left\{(\theta + 1)/\theta\right\}$. By specifying functional forms for the distribution parameters $\eta(X, t)$ and $\theta(t)$, a class of models for nonsymmetric distributions with analytic conditional means was obtained. These models are conditionally heteroskedastic with variance $\eta(X, t)^2\left(\Gamma\left[\{\theta(t) + 2\}/\theta(t)\right] - \left\{\Gamma\left[\{\theta(t) + 2\}/\theta(t)\right]\right\}2\right)$.

As in Cattaneo *et al.* [35], we generated Y_i conditional on X_{1i}, X_{2i} and T_i with $\theta_i = T_i + 1$ and $\eta_i = (\theta_i/3)(2 + X_{1i} + X_{2i} + X_{1i}^2 + X_{2i}^2 + X_{1i}X_{2i})$. Thus, the regression of Y_i on X_{1i}, X_{1i}^2, X_{2i}, X_{2i}^2 and $X_{1i} \times X_{2i}$

corresponds to the correctly specified model, whereas the regression of Y_i only on X_{1i} and X_{1i}^2 corresponds to the misspecified model. A covariate selection process was not implemented with these models in order to maintain the fidelity of the correct and incorrect specification.

4.3. *Model estimation*

In this section, we describe the estimation and inference procedures for each model and repetition over the four scenarios and two sample sizes. All simulations and analyses reported in this paper were conducted using Stata version 13.0 (College Station, TX, USA) [37].

For each of the four scenarios, six different methods were used to estimate the potential outcome mean for each of the three treatment levels. (i) Naïve parameter estimates were derived by regressing the outcome Y on indicator variables representing the levels of T. (ii) The RA estimator was implemented using the `teffects ra` command as described in Section 3.1. (iii) The IPTW estimator with adjusted weights was implemented using the `teffects ipw` command. (iv) MMWS estimates were derived by dividing the sample equally into ten strata based on the estimated GPS, computing the MMWS weights by implementing a user-written command for Stata `mmws` [38], and then by regressing the outcome Y on indicator variables representing the levels of T, with the MMWS weights used as sampling weights and applying robust standard errors. (v) The A-IPTW estimator was implemented using the `teffects aipw` command, which corresponds to Cattaneo's efficient influence function estimator [17]. (vi) The IPTW-RA estimator was implemented using the `teffects ipwra` command. All Stata `teffects` commands use a one-step GMM approach to provide correct standard errors of treatment effects taking into account estimated GPS [33]. After each model was estimated, we tested the hypothesis that the estimated coefficient was equal to the true value (for each given treatment level/scenario) using a Wald test. We then generated a dummy variable equaling 1 when the P value < 0.05. Summing these provided the false rejection rate–that is, the probability of committing a type I error.

4.4. *Monte Carlo simulation results*

Tables I–IV provide detailed results of the Monte Carlo simulations for each of the four scenarios. Each table is laid out as follows: the first 8 columns of the table provide results when $N = 500$, and the second 8 columns of the table provide results when $N = 2,000$, in the same order. The first column indicates the estimator, the second column provides the mean of the point estimates over the 10,000 repetitions, the third column provides the percent relative bias in the mean point estimates, the fourth column provides the mean squared error (MSE), the fifth column provides the standard deviation of the point estimates over the 10,000 repetitions, the sixth column provides the mean of the estimated standard errors over the 10,000 repetitions, the seventh column provides the ratio of the standard deviation of the point estimates to the mean of the estimated standard errors (SD/SE ratio), and the eighth column provides the mean of the rejection indicators over the 10,000 repetitions. The standard deviation of the point estimates should be as close as possible to the mean of the estimated standard errors resulting in a SD/SE ratio close to 1.0, and the mean of the false rejection indicators should be 0.05. These metrics can be considered as either of the following: (i) indicators of accuracy for the point estimates (i.e., bias, MSE), or (ii) indicators of accuracy in the distributions and inference (i.e., SD/SE ratio, rejection rate).

When both the treatment model (estimating the GPS) and outcome model were correctly specified (Table I), all estimation methods derived similarly accurate results for point estimates, bias, and MSE, across all treatment levels. The SD/SE ratio and rejection rate were similar across all methods with the exception of MMWS, which was consistently lower than all other methods. All adjustment methods were superior to the naïve estimates. In the case when both models were correctly specified, the consistency of all methods becomes clear when the MSEs over the two different sample sizes are compared. As expected, the ratio of MSEs for the two sample sizes were similar.

When the treatment model was correctly specified and the outcome model was misspecified (Table II), regression adjustment always provided mean point estimates that were furthest from the true values, together with the highest bias and MSE, across all treatment levels. All other adjustment methods provided point estimates close to the true values with nearly identical MSE. The MMWS estimator had lower SD/SE ratios and mean rejection rates than the other estimators, while RA consistently had much higher rejection rates. All adjustment methods were superior to the naïve estimates. The smaller change in the MSE for the RA estimates, even as the samples size increased, indicates that RA will not converge to the true value if the outcome model is misspecified.

Table I. Monte Carlo results for estimators when both the generalized propensity score and outcome models are correctly specified.

	N = 500							N = 2000						
Estimator	Mean estimate[1]	Relative bias (%)	MSE	SD of estimates	Mean SE	SD/SE ratio	Rejection rate	Mean estimates[1]	Relative bias (%)	MSE	SD of estimates	Mean SE	SD/SE ratio	Rejection rate
$T = 0$ ($\mu_0 = 0.7222$)														
Naïve	0.6989	−3.22	0.4460	0.0626	0.0665	0.9409	0.0508	0.6803	-5.81	0.2380	0.0249	0.0275	0.9058	0.3259
RA	0.7230	0.11	0.4756	0.0690	0.0661	1.0433	0.0684	0.7223	0.02	0.0905	0.0301	0.0299	1.0051	0.0530
MMWS	0.7217	−0.07	0.4876	0.0698	0.0695	1.0053	0.0631	0.7212	-0.14	0.0902	0.0300	0.0305	0.9852	0.0503
IPTW	0.7206	−0.22	0.4546	0.0674	0.0658	1.0237	0.0660	0.7223	0.02	0.0919	0.0303	0.0302	1.0033	0.0528
A-IPTW	0.7230	0.11	0.4759	0.0690	0.0661	1.0435	0.0683	0.7223	0.02	0.0910	0.0302	0.0300	1.0053	0.0533
IPTW-RA	0.7230	0.11	0.4781	0.0691	0.0660	1.0470	0.0694	0.7223	0.02	0.0910	0.0302	0.0300	1.0061	0.0540
$T = 1$ ($\mu_1 = 1.2801$)														
Naïve	1.2786	−0.11	0.3094	0.0556	0.0655	0.8486	0.0205	1.3439	4.98	0.5087	0.0319	0.0368	0.8678	0.3892
RA	1.2808	0.05	0.2924	0.0541	0.0536	1.0096	0.0536	1.2804	0.02	0.0846	0.0291	0.0289	1.0056	0.0533
MMWS	1.2805	0.03	0.3009	0.0549	0.0572	0.9598	0.0444	1.2814	0.10	0.0856	0.0292	0.0305	0.9583	0.0421
IPTW	1.2807	0.05	0.2920	0.0540	0.0538	1.0036	0.0527	1.2804	0.02	0.0844	0.0290	0.0289	1.0047	0.0517
A-IPTW	1.2808	0.05	0.2924	0.0541	0.0536	1.0096	0.0537	1.2804	0.02	0.0846	0.0291	0.0289	1.0059	0.0527
IPTW-RA	1.2808	0.05	0.2924	0.0541	0.0535	1.0099	0.0534	1.2804	0.02	0.0846	0.0291	0.0289	1.0069	0.0522
$T = 2$ ($\mu_2 = 1.9348$)														
Naïve	1.9820	2.44	0.5847	0.0601	0.0609	0.9874	0.1109	1.9859	2.64	0.3688	0.0329	0.0341	0.9651	0.3196
RA	1.9347	−0.01	0.2890	0.0538	0.0531	1.0120	0.0521	1.9353	0.03	0.0852	0.0292	0.0288	1.0130	0.0540
MMWS	1.9362	0.07	0.2972	0.0545	0.0580	0.9395	0.0383	1.9359	0.06	0.0861	0.0293	0.0314	0.9338	0.0340
IPTW	1.9359	0.06	0.2893	0.0538	0.0533	1.0083	0.0526	1.9353	0.02	0.0850	0.0292	0.0289	1.0102	0.0543
A-IPTW	1.9347	−0.01	0.2890	0.0538	0.0531	1.0120	0.0521	1.9353	0.03	0.0851	0.0292	0.0288	1.0130	0.0535
IPTW-RA	1.9347	−0.01	0.2889	0.0537	0.0531	1.0125	0.0523	1.9353	0.03	0.0851	0.0292	0.0288	1.0129	0.0543

Note: [1] Mean estimates represent potential outcome means. MMWS is marginal mean weighting through stratification, IPTW is inverse probability of treatment weighting, A-IPTW is augmented inverse probability of treatment weighting, and IPTW-RA is inverse probability of treatment weighting with regression adjustment, SD is standard deviation and SE is standard error.

Table II. Monte Carlo results for estimators when the generalized propensity score is correctly specified and outcome model is misspecified.

	N = 500							N = 2000						
Estimator	Mean estimate[1]	Relative bias (%)	MSE estimates	SD of SE	Mean ratio	SD/SE rate	Rejection estimates[1]	Mean bias (%)	Relative estimates	MSE SE	SD of ratio	Mean rate	SD/SE	Rejection
$T = 0$ ($\mu_0 = 0.7222$)														
Naïve	0.6994	−3.15	0.4397	0.0623	0.0665	0.9367	0.0503	0.6799	−5.86	0.2427	0.0252	0.0275	0.9175	0.3293
Regression adjustment	0.7102	−1.66	0.4350	0.0649	0.0636	1.0199	0.0774	0.699	−3.21	0.1289	0.0274	0.0273	1.0045	0.1593
MMWS	0.7221	−0.02	0.4829	0.0695	0.0693	1.0034	0.0622	0.7207	−0.21	0.0915	0.0302	0.0305	0.9913	0.0544
IPTW	0.7211	−0.15	0.4520	0.0672	0.0657	1.0229	0.0658	0.7218	−0.05	0.0929	0.0305	0.0302	1.0088	0.0557
A-IPTW	0.7216	−0.09	0.4552	0.0675	0.0657	1.0272	0.0668	0.7218	−0.06	0.0923	0.0304	0.0301	1.0088	0.0557
IPTW-RA	0.7216	−0.09	0.4556	0.0675	0.0657	1.0280	0.0674	0.7218	−0.06	0.0922	0.0304	0.0301	1.0092	0.0556
$T = 1$ ($\mu_1 = 1.2801$)														
Naïve	1.2775	−0.20	0.3130	0.0559	0.0655	0.8529	0.0210	1.3434	4.94	0.5020	0.0319	0.0368	0.8666	0.3904
Regression adjustment	1.2786	−0.12	0.3046	0.0552	0.0546	1.0102	0.0553	1.3111	2.42	0.1877	0.0302	0.0304	0.9953	0.1693
MMWS	1.2799	−0.02	0.3069	0.0554	0.0572	0.9690	0.0458	1.2811	0.08	0.0850	0.0291	0.0305	0.9560	0.0432
IPTW	1.2798	−0.02	0.2958	0.0544	0.0538	1.0107	0.0552	1.2801	0	0.0838	0.0290	0.0289	1.0018	0.0535
A-IPTW	1.2799	−0.02	0.2959	0.0544	0.0537	1.0124	0.0566	1.2802	0.01	0.0838	0.0290	0.0289	1.0015	0.0529
IPTW-RA	1.2799	−0.02	0.2960	0.0544	0.0537	1.0126	0.0560	1.2802	0.01	0.0838	0.0289	0.0289	1.0018	0.0531
$T = 2$ ($\mu_2 = 1.9348$)														
Naïve	1.9823	2.45	0.5913	0.0605	0.0609	0.9936	0.1163	1.9857	2.63	0.3643	0.0324	0.0340	0.9520	0.3088
Regression adjustment	1.9589	1.25	0.3874	0.0574	0.0563	1.0192	0.0690	1.9597	1.29	0.1552	0.0306	0.0306	0.9999	0.1251
MMWS	1.9368	0.11	0.3033	0.0550	0.0580	0.9489	0.0371	1.9356	0.04	0.0847	0.0291	0.0314	0.9261	0.0356
IPTW	1.9367	0.10	0.2949	0.0543	0.0533	1.0174	0.0552	1.935	0.01	0.0831	0.0288	0.0289	0.9982	0.0496
A-IPTW	1.9368	0.11	0.2961	0.0544	0.0534	1.0189	0.0558	1.9351	0.02	0.0832	0.0288	0.0289	0.9988	0.0508
IPTW-RA	1.9368	0.11	0.2960	0.0544	0.0533	1.0191	0.0554	1.9351	0.02	0.0832	0.0288	0.0289	0.9989	0.0505

Note: [1]Mean estimates represent potential outcome means. MMWS is marginal mean weighting through stratification, IPTW is inverse probability of treatment weighting, A-IPTW is augmented inverse probability of treatment weighting, and IPTW-RA is inverse probability of treatment weighting with regression adjustment, SD is standard deviation and SE is standard error.

Table III. Monte Carlo results for estimators when the generalized propensity score is misspecified, and the outcome model is correctly specified.

	N = 500							N = 2000						
Estimator	Mean estimate[1]	Relative bias (%)	MSE	SD of estimates	Mean SE	SD/SE ratio	Rejection rate	Mean estimates[1]	Relative bias (%)	MSE	SD of estimates	Mean SE	SD/SE ratio	Rejection rate
$T = 0$ ($\mu_0 = 0.7222$)														
Naïve	0.6982	−3.32	0.4557	0.0631	0.0665	0.9489	0.0550	0.6805	−5.77	0.2385	0.0254	0.0275	0.9235	0.3191
Regression adjustment	0.7224	0.03	0.4909	0.0701	0.0659	1.0635	0.0738	0.7225	0.05	0.0926	0.0304	0.0300	1.0157	0.0573
MMWS	0.7092	−1.81	0.4699	0.0673	0.0660	1.0202	0.0786	0.6994	−3.16	0.1287	0.0277	0.0275	1.0039	0.1495
IPTW	0.7089	−1.84	0.4482	0.0656	0.0636	1.0322	0.0821	0.6997	−3.12	0.1272	0.0276	0.0273	1.0110	0.1507
A-IPTW	0.7224	0.03	0.4909	0.0701	0.0659	1.0636	0.0739	0.7226	0.05	0.0926	0.0304	0.0300	1.0158	0.0575
IPTW-RA	0.7224	0.03	0.4926	0.0702	0.0659	1.0658	0.0742	0.7226	0.05	0.0931	0.0305	0.0300	1.0169	0.0574
$T = 1$ ($\mu_1 = 1.2801$)														
Naïve	1.2786	−0.11	0.3152	0.0561	0.0654	0.8581	0.0235	1.3439	4.98	0.5096	0.0321	0.0368	0.8724	0.3887
Regression adjustment	1.2808	0.05	0.3003	0.0548	0.0535	1.0241	0.0592	1.2803	0.01	0.0840	0.0290	0.0289	1.0025	0.0499
MMWS	1.2797	−0.03	0.3170	0.0563	0.0568	0.9904	0.0498	1.3118	2.47	0.1945	0.0307	0.0313	0.9821	0.1603
IPTW	1.2796	−0.04	0.3064	0.0554	0.0547	1.0128	0.0553	1.3115	2.45	0.1915	0.0305	0.0304	1.0043	0.1691
A-IPTW	1.2808	0.05	0.3002	0.0548	0.0535	1.0240	0.0592	1.2803	0.01	0.0840	0.0290	0.0289	1.0025	0.0498
IPTW-RA	1.2808	0.05	0.3003	0.0548	0.0535	1.0243	0.0590	1.2803	0.01	0.0839	0.0290	0.0289	1.0027	0.0499
$T = 2$ ($\mu_2 = 1.9348$)														
Naïve	1.9818	2.43	0.5780	0.0597	0.0609	0.9819	0.1137	1.9854	2.62	0.3610	0.0324	0.0341	0.9499	0.3073
Regression adjustment	1.9348	0	0.2845	0.0533	0.0531	1.0042	0.0507	1.9349	0.01	0.0828	0.0288	0.0288	0.9992	0.0507
MMWS	1.9584	1.22	0.3841	0.0573	0.0592	0.9677	0.0576	1.9596	1.28	0.1553	0.0306	0.0318	0.9627	0.1058
IPTW	1.9584	1.22	0.3746	0.0564	0.0564	1.0017	0.0676	1.9594	1.27	0.1534	0.0305	0.0306	0.9976	0.1226
A-IPTW	1.9348	0	0.2845	0.0533	0.0531	1.0042	0.0508	1.9349	0.01	0.0828	0.0288	0.0288	0.9992	0.0507
IPTW-RA	1.9348	0	0.2844	0.0533	0.0531	1.0044	0.0509	1.9349	0.01	0.0828	0.0288	0.0288	0.9992	0.0510

Note: [1]Mean estimates represent potential outcome means. MMWS is marginal mean weighting through stratification, IPTW is inverse probability of treatment weighting, A-IPTW is augmented inverse probability of treatment weighting, and IPTW-RA is inverse probability of treatment weighting with regression adjustment, SD is standard deviation and SE is standard error.

Table IV. Monte Carlo results for estimators when both the generalized propensity score and the outcome models are misspecified.

	N = 500							N = 2000						
Estimator	Mean estimate[1]	Relative bias (%)	MSE	SD of estimates	Mean SE	SD/SE ratio	Rejection rate	Mean estimates[1]	Relative bias (%)	MSE	SD of estimates	Mean SE	SD/SE ratio	Rejection rate
$T = 0$ ($\mu_0 = 0.7222$)														
Naïve	0.6996	−3.13	0.4398	0.0623	0.0665	0.9373	0.0507	0.6802	-5.82	0.2379	0.0247	0.0275	0.8993	0.3247
Regression adjustment	0.7103	−1.65	0.4348	0.0648	0.0637	1.0179	0.0769	0.6994	-3.16	0.1242	0.0269	0.0273	0.9853	0.1496
MMWS	0.7100	−1.69	0.4578	0.0666	0.0660	1.0085	0.0735	0.6992	-3.19	0.1256	0.0269	0.0275	0.9786	0.1487
IPTW	0.7099	−1.71	0.4320	0.0646	0.0637	1.0139	0.0748	0.6994	-3.16	0.1245	0.0269	0.0273	0.9852	0.1493
A-IPTW	0.7103	−1.65	0.4348	0.0648	0.0637	1.0179	0.0769	0.6994	-3.16	0.1243	0.0269	0.0273	0.9854	0.1498
IPTW-RA	0.7103	−1.65	0.4351	0.0649	0.0637	1.0183	0.0768	0.6994	-3.16	0.1242	0.0269	0.0273	0.9854	0.1499
$T = 1$ ($\mu_1 = 1.2801$)														
Naïve	1.2780	−0.17	0.3181	0.0564	0.0655	0.8600	0.0236	1.3435	4.95	0.5063	0.0323	0.0368	0.8767	0.3862
Regression adjustment	1.2790	−0.09	0.3082	0.0555	0.0546	1.0161	0.0579	1.3112	2.43	0.1903	0.0306	0.0304	1.0078	0.1705
MMWS	1.2789	−0.10	0.3185	0.0564	0.0569	0.9923	0.0532	1.3114	2.45	0.1923	0.0307	0.0312	0.9825	0.1603
IPTW	1.2790	−0.08	0.3072	0.0554	0.0547	1.0129	0.0582	1.3112	2.43	0.1902	0.0306	0.0304	1.0073	0.1708
A-IPTW	1.2790	−0.09	0.3082	0.0555	0.0546	1.0161	0.0579	1.3112	2.43	0.1903	0.0306	0.0304	1.0077	0.1706
IPTW-RA	1.2790	−0.09	0.3082	0.0555	0.0546	1.0161	0.0580	1.3112	2.43	0.1903	0.0306	0.0304	1.0079	0.1710
$T = 2$ ($\mu_2 = 1.9348$)														
Naïve	1.9828	2.48	0.5916	0.0601	0.0609	0.9859	0.1108	1.9851	2.60	0.3585	0.0325	0.0340	0.9549	0.2998
Regression adjustment	1.9592	1.26	0.3814	0.0567	0.0563	1.0078	0.0699	1.959	1.25	0.1526	0.0306	0.0305	1.0035	0.1223
MMWS	1.9595	1.28	0.3892	0.0573	0.0592	0.9669	0.0574	1.9593	1.27	0.1549	0.0308	0.0318	0.9683	0.1060
IPTW	1.9596	1.28	0.3834	0.0567	0.0564	1.0065	0.0695	1.959	1.25	0.1529	0.0307	0.0306	1.0037	0.1213
A-IPTW	1.9592	1.26	0.3814	0.0567	0.0563	1.0078	0.0698	1.959	1.25	0.1526	0.0306	0.0305	1.0035	0.1221
IPTW-RA	1.9592	1.26	0.3813	0.0567	0.0563	1.0079	0.0698	1.959	1.25	0.1526	0.0306	0.0305	1.0035	0.1220

Note: [1]Mean estimates represent potential outcome means. MMWS is marginal mean weighting through stratification, IPTW is inverse probability of treatment weighting, A-IPTW is augmented inverse probability of treatment weighting, and IPTW-RA is inverse probability of treatment weighting with regression adjustment, SD is standard deviation and SE is standard error.

When the treatment model was misspecified and the outcome model was correctly specified (Table III), MMWS and IPTW provided point estimates that were far from the true values and performed poorly on all accuracy metrics. Conversely, RA and the two doubly robust methods (A-IPTW and IPTW-RA) provided point estimates very close to the true value and were close to the ideal on all other accuracy metrics. All adjustment methods were superior to the naïve estimates. In comparing the MSEs across the two samples sizes, we see that MMWS and IPTW have lower proportional change as the sample size increases, indicating that even for smaller sample sizes these methods do not converge to the true values.

When both the treatment and outcome models were misspecified (Table IV), all estimators performed poorly and provided nearly identical poor estimates for all accuracy measures across all treatment levels. Even so, all adjustment methods were still superior to the naïve estimates. In comparing the MSEs across the two samples sizes, it is evident that none of the estimators were consistent when both models were misspecified.

5. Empirical example

5.1. Data

Our data come from a disease management program designed for patients with congestive heart failure and implemented in a large health plan located in the western United States. Individuals with the condition were contacted and invited to enroll in the program. Those agreeing to participate received one of the following interventions: (i) periodic telephone calls from a nurse to discuss self-management behaviors, or (ii) remote tele-monitoring (RTM), which entailed daily electronic transmission of the participant's disease-related symptoms to a database followed by a call from the nurse if symptoms appeared to indicate the onset of an acute exacerbation. Assignment to either intervention arm was conducted by the program nurse and based largely on a subjective assessment of the patient's psycho-social needs, past levels of health care utilization, and the patient's preferred level of contact. The primary goal of the intervention was to reduce avoidable hospitalizations [39]. Patients with congestive heart failure, but not participating in the program, received their usual medical care and served as controls in this study. We use these data solely to compare the treatment effect estimators, and our analyses do not represent a definitive assessment of the program's effectiveness.

The retrospectively collected data consist of observations for 1359 program participants who completed a full 12 months of the intervention and 6612 non-participants who were health-plan members during the same period but were not exposed to the intervention. The sample was divided according to treatment assignment: (i) 6612 non-participants, (ii) 654 participants in the telephonic intervention, and (iii) 705 participants in the RTM intervention. Each individual in the dataset has 12 months of pre-intervention data and 12 months of intervention-period data. Pre-intervention characteristics of participants in the three study arms include patient demographic characteristics (age and gender), the Charlson comorbidity index and associated comorbidities [40], and key measures of health care utilization (prescription filled, office visits, emergency department visits, hospital admissions and hospital days). The primary outcome for all analyses used in this paper is the number of all-cause hospitalizations in the intervention year.

5.2. Evaluation approach

All analyses were performed using the same methods described in Section 3 but replacing ordinary least squares with Poisson regression for the outcome model. The GPS was estimated using multinomial logistic regression with the three-level treatment variable as the outcome. The choice of right-side variables was determined using `bfit` in Stata as described in Section 3. To implement the MMWS estimator, the study sample was divided equally into five strata based on the estimated GPS. For each estimator, potential outcome means were estimated for each treatment level. Additionally, pairwise contrasts (treatment effects) were estimated between all treatment levels, and across all estimators studied, using Stata's `pwcompare` command. `pwcompare` performs Wald tests using linear combinations of marginal linear predictions and uses the delta method to estimate the variance. *P* values are then Bonferroni adjusted to account for multiple comparisons.

5.3. Empirical results

Figures 1–3 show the kernel density estimates for the GPS for each treatment arm. In none of the plots does there appear to be a probability mass near 0 or 1, and the three estimated densities have most of

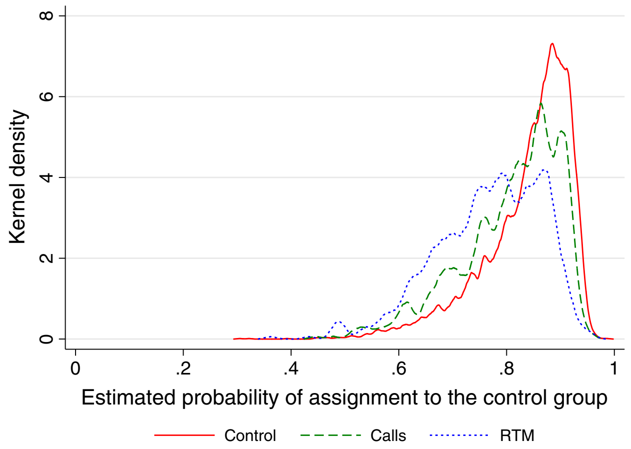

Figure 1. Overlap graph of the probability of assignment to the control condition. The density of the probability of assignment to the control group is estimated by non-parametric kernel density estimation with a triangular kernel and optimal bandwidth chosen by Stata and is plotted by treatment group. RTM, remote telemonitoring.

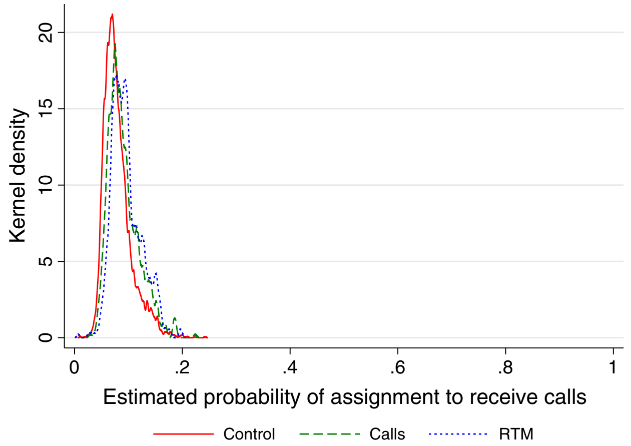

Figure 2. Overlap graph of the probability of assignment to receive calls. The density of the probability of assignment to the control group is estimated by non-parametric kernel density estimation with a triangular kernel and optimal bandwidth chosen by Stata, and is plotted by treatment group. RTM, remote telemonitoring.

their respective masses in regions in which they overlap each other. Therefore, there is no evidence that the overlap assumption is violated.

Table V presents the unadjusted pre-intervention characteristics of participants in the three study arms. The absolute standardized mean difference (SMD) is used to assess covariate balance [26]. Many of the SMDs are substantially greater than zero (optimal), and nine of the 69 SMDs are greater than the 0.25 cut-off recommended by Rubin [41]. In general, the participants in the RTM intervention arm were older and had a higher prevalence of comorbidities than the other two groups. However, all groups were comparable on key measures of health care utilization. Table VI presents the same pre-intervention characteristics of the study participants after weighting. As shown, all SMDs are much closer to zero, and no value is greater than 0.25.

Table VII provides the potential outcome means for each treatment, by estimator. All adjusted models offer nearly identical estimates for the control group's hospital admissions and provide similar estimates for the other two treatment arms. Interestingly, the naïve model provides a lower point estimate for the control group than the adjusted methods and higher point estimates for the two intervention arms as compared with the adjusted methods, indicating a downward bias because of omitted variables for the former and upward bias for latter.

Table VIII provides pairwise treatment effect estimates between all treatment arms, by estimator. Here, treatment effects represent the difference between groups in all cause hospital admissions. The naïve model showed that both intervention arms (calls and RTM) had significantly higher rates of hospital

Table V. Unadjusted baseline (prior 12 months) characteristics of program participants and non-participants in a multivalued treatment study.

				Absolute Standardized Differences		
Variables	Control	Calls	RTM	Calls vs Controls	RTM vs Controls	RTM vs Calls
N	6612	654	705			
Female	2976 (45.0%)	308 (47.1%)	343 (48.7%)	0.042	0.073	0.031
Age, mean (SD)	62.96 (15.77)	66.17 (14.55)	72.31 (12.42)	0.212	0.659	0.454
Charlson Comorbidity Index Score, mean (SD)	2.64 (2.52)	3.28 (2.57)	3.67 (2.63)	0.250	0.399	0.150
Diabetes - non complicated	1723 (26.1%)	244 (37.3%)	281 (39.9%)	0.244	0.297	0.052
Diabetes - complicated	697 (10.5%)	122 (18.7%)	130 (18.4%)	0.231	0.226	0.006
Acute myocardial infarction	782 (11.8%)	112 (17.1%)	162 (23.0%)	0.151	0.297	0.147
Chronic lung disease	1468 (22.2%)	177 (27.1%)	251 (35.6%)	0.113	0.299	0.185
Liver disease - mild	396 (6.0%)	35 (5.4%)	32 (4.5%)	0.028	0.065	0.038
Liver disease - moderate/ severe	48 (0.7%)	3 (0.5%)	5 (0.7%)	0.035	0.002	0.033
Cancer	784 (11.9%)	80 (12.2%)	97 (13.8%)	0.012	0.057	0.045
Cancer - metastic	140 (2.1%)	11 (1.7%)	10 (1.4%)	0.032	0.053	0.021
Rheumatoid Disease	228 (3.4%)	30 (4.6%)	26 (3.7%)	0.058	0.013	0.045
Cerebrovascular disease	952 (14.4%)	107 (16.4%)	132 (18.7%)	0.054	0.117	0.062
Peripheral vascular disease	874 (13.2%)	115 (17.6%)	150 (21.3%)	0.121	0.214	0.093
Renal disease	1083 (16.4%)	160 (24.5%)	214 (30.4%)	0.202	0.335	0.132
Dementia	164 (2.5%)	10 (1.5%)	8 (1.1%)	0.068	0.101	0.034
Hemiplegia or Paraplegia	130 (2.0%)	10 (1.5%)	11 (1.6%)	0.033	0.031	0.003
Peptic ulcer disease	105 (1.6%)	12 (1.8%)	13 (1.8%)	0.019	0.020	0.001
Prescriptions, mean (SD)	41.10 (37.42)	49.37 (38.90)	55.32 (37.18)	0.217	0.381	0.156
Office visits, mean (SD)	0.42 (0.93)	0.47 (0.83)	0.44 (0.84)	0.056	0.014	0.044
Emergency department visits, mean (SD)	0.49 (1.30)	0.51 (1.04)	0.44 (0.95)	0.017	0.046	0.072
Baseline hospital admissions, mean (SD)	0.64 (1.15)	0.74 (1.07)	0.64 (1.04)	0.088	0.006	0.099
Hospital days, mean (SD)	3.66 (11.61)	3.74 (8.60)	3.21 (16.09)	0.008	0.032	0.041

Note: [1]All variables are reported as N (%) unless otherwise noted. [2] RTM, remote telemonitoring.

Table VI. Weighted baseline (prior 12 months) characteristics of program participants and non-participants in a multivalued treatment study.

				Absolute Standardized Differences		
Variables	Control	Calls	RTM	Calls vs Controls	RTM vs Controls	RTM vs Calls
N	6612	654	705			
Female	1227 (45.5)	1227 (45.9)	1182 (45.5)	0.007	0.001	0.008
Age, mean (SD)	64.11 (15.85)	64.50 (15.05)	65.72 (13.89)	0.026	0.109	0.086
Charlson Comorbidity Index Score, mean (SD)	2.79 (2.59)	2.83 (2.43)	2.90 (2.51)	0.014	0.044	0.029
Diabetes - non complicated	764 (28.3%)	792 (29.6%)	757 (29.1%)	0.028	0.018	0.010
Diabetes - complicated	323 (12.0%)	332 (12.4%)	343 (13.2%)	0.013	0.037	0.021
Acute myocardial infarction	359 (13.3%)	368 (13.8%)	352 (13.6%)	0.014	0.008	0.006
Chronic lung disease	642 (23.8%)	638 (23.9%)	634 (24.4%)	0.001	0.014	0.013
Liver disease - mild	157 (5.8%)	164 (6.1%)	147 (5.7%)	0.013	0.007	0.021
Liver disease - moderate/ severe	19 (0.7%)	18 (0.7%)	22 (0.8%)	0.002	0.016	0.020
Cancer	325 (12.1%)	302 (11.3%)	312 (12.0%)	0.024	0.001	0.022
Cancer - metastic	54 (2.0%)	52 (1.9%)	50 (1.9%)	0.006	0.007	0.002
Rheumatoid Disease	95 (3.5%)	92 (3.4%)	96 (3.7%)	0.006	0.008	0.013
Cerebrovascular disease	405 (15.0%)	426 (15.9%)	444 (17.1%)	0.024	0.056	0.031
Peripheral vascular disease	388 (14.4%)	408 (15.3%)	396 (15.2%)	0.025	0.024	0.001
Renal disease	495 (18.4%)	493 (18.4%)	513 (19.7%)	0.002	0.035	0.031
Dementia	62 (2.3%)	70 (2.6%)	87 (3.3%)	0.022	0.063	0.046
Hemiplegia or Paraplegia	51 (1.9%)	56 (2.1%)	81 (3.1%)	0.014	0.078	0.068
Peptic ulcer disease	44 (1.6%)	42 (1.6%)	47 (1.8%)	0.007	0.013	0.019
Prescriptions, mean (SD)	43.24 (39.07)	43.92 (36.00)	46.22 (35.14)	0.019	0.082	0.062
Office visits, mean (SD)	0.43 (0.92)	0.44 (0.82)	0.44 (0.84)	0.015	0.012	0.003
Emergency department visits, mean (SD)	0.49 (1.26)	0.50 (1.07)	0.47 (0.98)	0.015	0.014	0.033
Baseline hospital admissions, mean (SD)	0.65 (1.15)	0.67 (1.02)	0.65 (1.07)	0.018	0.000	0.019
Hospital days, mean (SD)	3.64 (11.20)	3.47 (8.63)	3.50 (18.13)	0.017	0.009	0.002

Note: [1]All variables are reported as N (%) unless otherwise noted. [2] RTM, remote telemonitoring.

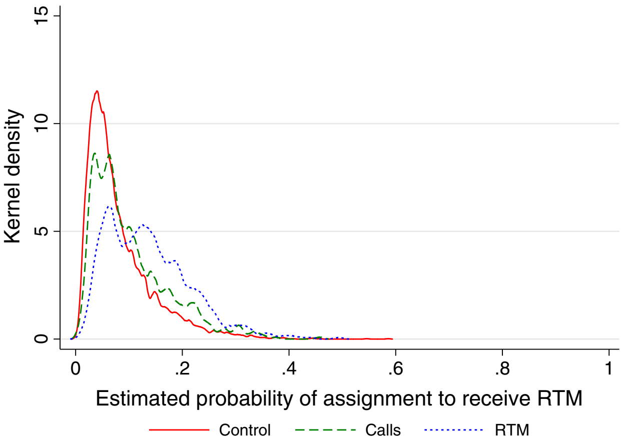

Figure 3. Overlap graph of the probability of assignment to receive remote telemonitoring. The density of the probability of assignment to the control group is estimated by non-parametric kernel density estimation with a triangular kernel and optimal bandwidth chosen by Stata, and is plotted by treatment group. RTM, remote telemonitoring.

Table VII. Potential outcome means for each treatment level by causal estimator.

Estimator	Mean estimate*	SE	z	*P* > |*z*|	[95% Conf. Interval]	
Control						
Naïve	0.315	0.011	28.69	<0.001	0.294	0.337
Regression adjustment	0.329	0.011	30.69	<0.001	0.308	0.350
MMWS	0.330	0.011	30.08	<0.001	0.309	0.352
IPTW	0.331	0.011	30.51	<0.001	0.310	0.352
A-IPTW	0.330	0.011	30.53	<0.001	0.309	0.351
IPTW-RA	0.330	0.011	30.54	<0.001	0.309	0.351
Calls						
Naïve	0.531	0.035	15.18	<0.001	0.462	0.599
Regression adjustment	0.508	0.045	11.38	<0.001	0.421	0.595
MMWS	0.523	0.051	10.2	<0.001	0.423	0.624
IPTW	0.512	0.042	12.07	<0.001	0.428	0.595
A-IPTW	0.509	0.044	11.67	<0.001	0.423	0.594
IPTW-RA	0.510	0.044	11.68	<0.001	0.424	0.595
RTM						
Naïve	0.444	0.034	13.19	<0.001	0.378	0.510
Regression adjustment	0.352	0.037	9.4	<0.001	0.279	0.425
MMWS	0.344	0.035	9.74	<0.001	0.275	0.413
IPTW	0.360	0.038	9.54	<0.001	0.286	0.434
A-IPTW	0.345	0.034	10.18	<0.001	0.279	0.412
IPTW-RA	0.344	0.032	10.79	<0.001	0.282	0.407

Note: *Estimates represent all-cause hospitalizations during the intervention period. MMWS is marginal mean weighting through stratification, IPTW is inverse probability of treatment weighting, A-IPTW is augmented inverse probability of treatment weighting, and IPTW-RA is inverse probability of treatment weighting with regression adjustment, RTM is remote telemonitoring.

admission than the control arm ($P < 0.001$ and $P = 0.001$ for calls versus controls, and RTM versus controls, respectively), but no statistically significant difference between the intervention arms themselves. All of the adjusted methods trended toward similar results. Irrespective of adjustment method, the arm receiving nursing calls had statistically higher hospital admissions than controls, the RTM arm was not statistically different than controls, and the RTM arm had statistically fewer admissions than the arm receiving nursing calls.

Table VIII. Pairwise treatment effects (Bonferroni adjusted) by causal estimator.

Estimator	Effect*	SE	z	$P > \|z\|$	[95% Conf.	Interval]
Naïve						
Calls vs Control	0.215	0.037	5.87	<0.001	0.127	0.303
RTM vs Control	0.128	0.035	3.63	0.001	0.044	0.213
RTM vs Calls	−0.087	0.049	−1.78	0.223	−0.203	0.030
Regression adjustment						
Calls vs Control	0.179	0.046	3.93	<0.001	0.070	0.288
RTM vs Control	0.023	0.039	0.59	1.000	−0.070	0.115
RTM vs Calls	−0.156	0.058	−2.69	0.021	−0.295	−0.017
MMWS						
Calls vs Control	0.193	0.052	3.68	0.001	0.067	0.319
RTM vs Control	0.013	0.037	0.36	1.000	−0.075	0.102
RTM vs Calls	−0.180	0.062	−2.88	0.012	−0.329	−0.030
IPTW						
Calls vs Control	0.180	0.043	4.16	<0.001	0.077	0.284
RTM vs Control	0.029	0.039	0.74	1.000	−0.064	0.122
RTM vs Calls	−0.152	0.057	−2.68	0.022	−0.287	−0.016
A-IPTW						
Calls vs Control	0.179	0.045	4.01	<0.001	0.072	0.285
RTM vs Control	0.015	0.035	0.44	1.000	−0.069	0.100
RTM vs Calls	−0.163	0.055	−2.97	0.009	−0.294	−0.032
IPTW-RA						
Calls vs Control	0.180	0.045	4.03	<0.001	0.073	0.286
RTM vs Control	0.014	0.033	0.43	1.000	−0.065	0.094
RTM vs Calls	−0.165	0.054	−3.09	0.006	−0.293	−0.037

6. Discussion

We used Monte Carlo simulations to compare the performance of several adjustment techniques with estimate treatment effects in multivalued treatments and empirical data to demonstrate the application of the methods. Our simulation results can be briefly summarized as follows: (i) when both the GPS model (estimating the treatment) and outcome model are correctly specified, all adjustment methods provide similarly unbiased estimates (ii) when the outcome model is misspecified, regression adjustment performs poorly, while all the weighting methods provide unbiased estimates (iii) when the GPS is misspecified, methods based solely on modeling the GPS (i.e., MMWS and IPTW) perform poorly, while RA and the doubly robust models provide unbiased estimates, and (iv) when both the GPS and outcome model are misspecified, all methods perform poorly.

Our finding that doubly robust methods consistently provide unbiased estimates when either the GPS or outcomes model is misspecified is supported by other simulation studies examining a similar array of adjustment methods applied to the binary treatment case [16, 20, 32, 42, 43] and in a recent study of multivalued treatments [19]. Thus, from a practical standpoint, the investigator may be best served by utilizing a doubly robust approach, as it is unlikely that he or she will be able to ascertain which of the two models is misspecified. There currently appears to be no consensus as to which approach is most appropriate if both models are misspecified [32, 43, 44].

Our results also indicate that the two doubly robust methods (A-IPTW and IPTW-RA) perform nearly identically under the different conditions imposed upon them in the simulations. This similarity extends to standard errors and interval coverage as well as bias. Therefore, investigators may rely on other considerations when choosing between methods. For example, A-IPTW, as implemented in Cattaneo *et al.* [35], can estimate quantile treatment effects, while IPTW-RA (as currently implemented in Stata) does not.

On the other hand, IPTW-RA can estimate average treatment effects on the treated, whereas the current A-IPTW method implemented in Stata does not (however a version recently written for R can estimate the average treatment effects on the treated [45]).

In our empirical example, we observed that all adjustment methods generally produced qualitatively similar results and all differed substantially from the naïve estimates. While any of the adjusted models would be an acceptable choice for analyzing these data, as mentioned above, utilizing a doubly robust approach will increase the chances of deriving an accurate treatment effects estimate when either the GPS or outcome model is misspecified. Thus, rather than estimating several models and hoping for a consistent result, a more economical approach would be to rely on doubly robust estimators as the principal evaluation approach.

Our study has limitations. First, we considered the performance of alternative multivalued treatment estimators in the context of a specific data generating process. It is unclear how estimator performance may vary across different data generating processes, such as heterogeneous treatments or with overlap problems. Second, our simulation assumed strong ignorability. By itself, this assumption limits much of the bias in observational studies (e.g., from confounding from unobservables). Thus, future research should explore the performance of multivalued treatment estimators in the context of more diverse data generating processes (including additional variable types and distributions), bootstrapping methods [46], and violations to assumptions in the causal model. In particular, we have no reason to assume the similarity of the methods would also hold for much smaller sample sizes where model parsimony might be relatively more important.

In conclusion, the results of our comprehensive simulation study suggest that investigators should consider doubly robust estimators as the principal evaluation approach in observational studies of multivalued treatments. Our findings are consistent with those reported for binary treatments. By supporting the use of doubly robust estimation for multivalued treatments, our findings extend previous results from binary treatment studies.

Acknowledgements

We thank the editor, Ralph D'Agostino, and three anonymous referees for their helpful reviews. We also wish to thank Julia Adler-Milstein for reviewing the manuscript and providing valuable comments and suggestions.

References

1. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**(3):706–710.
2. Lechner Michael. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, Lechner M, Pfeiffer F (eds). Physica: Heidelberg, 2001; 43–58.
3. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.
4. Lu B, Zanutto E, Hornik R, Rosenbaum PR. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* 2001; **96**:1245–1253.
5. Lechner M. Program heterogeneity and propensity score matching: an application to the evaluation of active labor market policies. *Review of Economics and Statistics* 2002; **84**(2):205–220.
6. Frölich M. Programme evaluation with multiple treatments. *Journal of Economic Surveys* 2004; **18**(2):181–224.
7. Blundell R, Dearden L, Sianesi B. Evaluating the impact of education on earnings in the UK: models, methods and results from the NCDS. *Journal of the Royal Statistical Society, Series A* 2005; **168**:473–512. IFS Working Papers W03/20, Institute for Fiscal Studies.
8. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data. 2nd ed.* MIT Press: Cambridge, MA, 2010.
9. Hong G. Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics* 2010; **35**:499–531.
10. Cattaneo MD, Farrell MH. Efficient estimation of the dose response function under ignorability using subclassification on the covariates. In *Missing Data Methods: Cross-Sectional Methods and Applications*, Drukker DM (ed.) Emerald Group Publishing Limited: Howard House, Wagon Lane, Bingley BD16 1WA, UK, 2011; 93–127.
11. Timbie JW, Fox DS, Van Busum K, Schneider EC. Five reasons that many comparative effectiveness studies fail to change patient care and clinical practice. *Health Affairs* 2012; **31**:2168–2175.
12. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987; **82**:387–394.
13. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 1995; **90**(429):122–129.
14. Hong G. Marginal mean weighting through stratification: a generalized method for evaluating multi-valued and multiple treatments with non-experimental data. *Psychological Methods* 2012; **17**:44–60.

15. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. In *1999 Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association: Alexandria, VA, 2000, 6–10.
16. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**(4): 962–973.
17. Cattaneo MD. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 2010; **155**(2):138 –154.
18. Wooldridge JM. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 2007; **141**(2):1281–1301.
19. Uysal SD. Doubly robust estimation of causal effects with multivalued treatments: an application to the returns to schooling. *Journal of Applied Econometrics* 2015; **30**(5):763–786.
20. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**:2937–2960.
21. Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.
22. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1):41–55.
23. Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 2009; **47**(1):5–86.
24. Drake C. Effects of misspecification of the propensity score on estimators of treatment effects. *Biometrics* 1993; **49**:1231–1236.
25. Horvitz DG, Thompson DJ, Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**(260):663–685.
26. Linden A, Samuels SJ. Using balance statistics to determine the optimal number of controls in matching studies. *Journal of Evaluation in Clinical Practice* 2013; **19**:968–975.
27. Linden A. Graphical displays for assessing covariate balance in matching studies. *Journal of Evaluation in Clinical Practice* 2015; **21**:242–247.
28. Kurth TA, Walker M, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* 2006; **163**:262–270.
29. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
30. Linden A. Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions. *Journal of Evaluation in Clinical Practice* 2014; **20**:1065–1071.
31. Huang IC, Frangakis C, Dominici F, Diette GB, Wu AW. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Services Research* 2005; **40**:253–278.
32. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 2007; **22**:523–539.
33. StataCorp. *Stata 13 Treatment Effects Manual: Potential Outcomes/Counterfactual Outcomes*. Stata Press: College Station, TX, 2013.
34. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology* 2001; **2**:259–278.
35. Cattaneo MD, Drukker DM, Holland A. Estimation of multivalued treatment effects under conditional independence. *Stata Journal* 2013; **13**:407–450.
36. Farrell MH. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* 2015; **189**:1–23.
37. StataCorp. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP, 2013.
38. Linden A. mmws: Stata module for implementing mean marginal weighting through stratification. Statistical Software Components s457886, Boston College Department of Economics, 2014. Downloadable from http://ideas.repec.org/c/boc/bocode/s457886.html [Accessed on 26 March 2015].
39. Linden A. What will it take for disease management to demonstrate a return on investment? New perspectives on an old theme. *American Journal of Managed Care* 2006; **12**:217–222.
40. Charlson ME, Pompei P, Ales KL, McKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Disease* 1987; **40**:373–383.
41. Rubin D. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2001; **2**(3-4):169–188.
42. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine* 2010; **29**:2137–2148.
43. Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* 2010; **97**(3):661–682.
44. Robins JM, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science* 2007; **22**(4):544–559.
45. Colonico S, Cattaneo MD. mvte: An R package for inference on multi-valued treatment effects, University of Michigan, 2015.
46. Linden A, Adams J, Roberts N. Evaluating disease management program effectiveness: an introduction to the bootstrap technique. *Disease Management and Health Outcomes* 2005; **13**:159–167.