Published in final edited form as:

Biometrics. 2010 September ; 66(3): 824–833. doi:10.1111/j.1541-0420.2009.01334.x.

Estimating treatment effects of longitudinal designs using regression models on propensity scores

Aristide C. Achy-Brou, Constantine E. Frangakis, and Michael Griswold*

Summary

We derive regression estimators that can compare longitudinal treatments using only the longitudinal propensity scores as regressors. These estimators, which assume knowledge of the variables used in the treatment assignment, are important for reducing the large dimension of covariates for two reasons. First, if the regression models on the longitudinal propensity scores are correct, then our estimators share advantages of correctly-specified model-based estimators, a benefit not shared by estimators based on weights alone. Second, if the models are incorrect, the misspecification can be more easily limited through model checking than with models based on the full covariates. Thus, our estimators can also be better when used in place of the regression on the full covariates. We use our methods to compare longitudinal treatments for type 2 diabetes mellitus.

Keywords

Admissibility; Longitudinal treatments; Observation study; Propensity scores

1. Introduction

We wish to estimate effects of longitudinal treatments in observational studies. Such estimation must control the growing dimension of variables that predict treatment assignment over time.

A motivating example is the comparison between sustained exposure to recently approved Exenatide versus insulin and various oral medications for type II diabetes, on monthly hospitalization frequency and monthly total health care charges outcomes. The retrospective study we use has three two-month periods, four choices of treatment to choose from for each period, three baseline covariates, 18 time-varying covariates and two univariate outcomes of interest.

One method for estimating the above effects is through Robins's (1987) g-computation formula. This method has rarely been used because it has been mostly known as needing to adjust for the entire covariate history. This is subject to a model misspecification that is

*Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA. cfrangak@jhsph.edu.

Supplementary Material

Web Appendices referenced in Sections 4–6 are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

difficult to check and fix. For this reason, the most widely used method for estimating the effect of sustained longitudinal treatments is derived from the Horvitz-Thompson (HT, 1952) inverse propensity score weighting approach using marginal structural models (Robins, Hernan and Brumback, 2000). This method provides generally consistent estimators when the propensity scores are correct. However, it can be inefficient, as standard methods for estimating parameters in these models do not use information on the time-varying covariates except through the weights.

In this paper, we derive regression estimators that can compare longitudinal treatments using only the longitudinal propensity scores as regressors. These estimators, which assume knowledge of the variables used in the treatment assignment, are important for reducing the large dimension of covariates, for two reasons. First, if the regression models are correct, then our estimators have the benefits of model-based estimators, benefits not shared by estimators based on weights alone. Second, like other estimators using models, our proposed estimator, requires more assumptions than estimators based on weights alone so is more subject to model misspecification. But the distance of our models from the truth can be more easily limited using model checking techniques, thanks to the reduced dimension of the regressors. Thus, by continuity, our estimators can also be better when used in place of the outcome regression on the full covariates, whether used alone or within augmented weighted estimators.

In Sec. 2., we briefly review the results for single time treatment (Rosenbaum and Rubin, 1983). In Sec. 3., we specify the assumptions and goals for our longitudinal setting. In Sec. 4., we provide two key results for using regression on propensity scores to estimate the effect of longitudinal treatments. We use a new method to estimate the effect of a new drug for treatment of type II diabetes mellitus in Sec. 5.. We perform model checks and provide a comparison with other methods. In Sec. 6., we discuss the role of our methods if the longitudinal models are misspecified. The Appendix provides technical details.

2. Review of single time propensity score regression

Consider comparing two treatments, labeled $Z = 1$ and 0 on an outcome Y in a simple random sample of patients from a population. For the i^{th} patient, denote by $Y_i(z)$ the potential outcomes that will be observed if the patient is assigned treatments $z = 0, 1$ (Neyman, 1923; Rubin, 1974; Rubin, 1978). Then, causal effects are comparisons of $Y_i(1)$ and $Y_i(0)$ for a common group of patients, for example $E\{Y(1) - Y(0)\}$.¹ Since each patient receives only one treatment at a particular time, either $Y_i(1)$ or $Y_i(0)$ is observed, but not both, so estimation of the effects requires certain assumptions.

For the i^{th} patient, let Z_i be the random variable indicating the treatment patient i could be assigned; let Z_i^{obs} be the actual treatment assigned: 1 for treatment, 0 otherwise; let $Y_i^{obs} = Y_i(Z_i^{obs})$ be the observed outcome; let X_i be a vector of observed pretreatment covariates; and denote the propensity score, the conditional probability of receiving

¹Here, expectations and probability statements refer to averages and frequencies among patients in the population of interest, rather than probabilities within a patient.

treatment $z = 1$, given the covariates, by $e(X) = \text{pr}(Z^{obs} = 1 | X)$. When there is thoughtful consideration of what covariates are measured in X_i , the following assumptions (Rubin, 1978) can be plausible:

Assumption (1') *Stable unit-treatment value assumption:* *The outcome of patient i to treatment z does not depend on the treatment that is to be assigned to patient j . This is implicit in the notation $Y_i()$ as function of only that patient's treatment.*

Assumption (2') *Ignorable treatment assignment:* *The covariates X_i have all the variables used in the treatment assignment mechanism. This implies that across patients, treatment is as if randomized with probabilities depending on the vector of observed covariates, i.e.,*

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp Z_i^{obs} | X_i$$

In theory X needs not include a variable used in the assignment mechanism but not related to the outcome. However, this would rarely be known a priori.

Assumption (3') *Random assignment :* *Every patient in our population has a positive probability of being exposed to treatment $Z^{obs} = 1$, $0 < e(X_i) < 1$.*

Under assumptions (1', 2', 3'), $E\{Y(z)\}$ can be expressed as $E_X \{E(Y^{obs} | X, Z^{obs} = z)\}$ for $z = 0, 1$. Thus the above causal effects can be expressed through subclassification (Rubin, 1974) on the covariates X ,

$$E\{Y(1)\} - E\{Y(0)\} = E_X \{E(Y^{obs} | X, Z^{obs} = 1)\} - E_X \{E(Y^{obs} | X, Z^{obs} = 0)\} \quad (1)$$

With a large sample, the regressions $E(Y^{obs} | X, Z^{obs} = 1)$ are estimable and so the causal effects $E\{Y(1) - Y(0)\}$ are estimable from (1). In practice, there may be too many covariates for positing checkable models $\text{pr}(Y^{obs} | X, Z^{obs} = 1)$. Such models are not necessary if one knows the propensity scores $e(X)$ (Rosenbaum and Rubin, 1983): they showed that the ignorability assumption (2') implies ignorability given only the propensity score, $(Y(1), Y(0)) \perp\!\!\!\perp Z^{obs} | e(X)$. Thus, in (1), we can use the propensity score $e(X)$ in place of X to get, for $z = 0, 1$

$$E\{Y(z)\} = E_e [E\{Y^{obs} | e(X), Z^{obs} = z\}] \quad (2)$$

Assuming the propensity score is known, (2) implies we can estimate the two causal effect in the LHS of (1) even if we keep only the propensity score as a summary of the covariates and estimate the expectation $E\{Y^{obs} | e(X), Z^{obs} = z\}$ and $\text{pr}\{e(X)\}$. In practice, $e(X)$ is estimated by eliciting information on the assignment mechanism from decision makers and using a model, e.g., logistic regression, on the elicited variables.

For multiple treatments, the use of the propensity scores as regressors have been discussed when these treatments are assigned ignorably at a single time (e.g., Imbens 2000). The design reflected by such an assumption is different from the design of *sequentially* ignorable

assigned treatments (Rubin, 1978; Robins 1987). Below we review the assumptions reflecting the latter design, which is the implicit or explicit design of many applications, including our own. For such a design, we derive in Sec. 4 regression estimators that compare longitudinal treatments using only the longitudinal propensity scores as regressors.

3. Assumptions and goals

Now suppose patients take treatments at multiple time, $t = 1, \dots, T$, and at each time t the treatment z_t can take values $1, \dots, K$, with $T \geq 2$ and $K \geq 2$. For the i^{th} patient and for a specified longitudinal treatment $z = (z_1, z_2, \dots, z_T)$, we denote by $Y_i(z) \equiv Y_i(z_1, z_2, \dots, z_T)$, the univariate outcome that would have been observed at the end of the study. For a particular longitudinal treatment z , we wish to estimate the distribution $\text{pr}(Y(z))$, and then compare among such distributions.

For the i^{th} patient, and at time t , denote the actual treatment assigned by $Z_{i,t}^{\text{obs}}$, let Z_i^{obs} be the vector of $Z_{i,t}^{\text{obs}}$, let $X_{i,t}^{\text{obs}}$ be the vector of variables observed after having been assigned to treatment at time $t - 1$ but before assignment of treatment at time t . Moreover, let the patient history $H_{i,t}$ be the cumulative information observed before the assignment at time t , that is,

$$H_{i,t} = \{(X_{i,1}^{\text{obs}}, Z_{i,1}^{\text{obs}}), \dots, (X_{i,t-1}^{\text{obs}}, Z_{i,t-1}^{\text{obs}}), X_{i,t}^{\text{obs}}\}.$$

Denote by Y_i^{obs} the observed outcome at the end of the last period, which, based on the potential outcomes notation, is equal to $Y_i(Z_i^{\text{obs}})$. Finally, let the conditional probability of receiving treatment $Z_{i,t}^{\text{obs}} = k$, given the history $H_{i,t}$, be denoted by

$$e_{i,t,k} = \text{pr}(Z_{i,t}^{\text{obs}} = k | H_{i,t}).$$

We make the following three assumptions, in analogy to Assumptions (1', 2', 3').

- Assumption (1)** *Stable unit-treatment value assumption:* The outcome of patient i to treatment $z = (z_1, \dots, z_t)$ does not depend on the treatment that is to be assigned to patient j . As with assumption (1'), this assumption is implicit in the notation $Y_i(z)$.
- Assumption (2)** *Sequential ignorable treatment assignment:* At every time t , the history $H_{i,t}$ has all the variables used in the mechanism of assigning treatment $Z_{i,t}^{\text{obs}}$. This implies that across patients,
 $\{Y_i(z) : \text{all } z\} \perp\!\!\!\perp Z_{i,t}^{\text{obs}} | H_{i,t}$, so at every time t the treatment is as if randomized with probabilities depending on the observed history.
- Assumption (3)** *Random assignment:* At each time point t , every patient in our population has a positive probability of being exposed to each treatment $Z_{i,t}^{\text{obs}} = k$, $0 < e_{i,t,k} < 1$.

Under the above assumptions, Robins (1987) showed that an induction on Rubin's (1974) technique that established (1) yields

$$\begin{aligned} pr(Y(z)) = & \sum_{X_1^{obs}, \dots, X_T^{obs}} pr(Y^{obs} | X_1^{obs}, Z_1^{obs} = z_1, \dots, X_T^{obs}, Z_T^{obs} = z_T) \times pr(X_1^{obs}) \\ & \times pr(X_2^{obs} | X_1^{obs}, Z_1^{obs} = z_1) \dots \times pr(X_T^{obs} | X_1^{obs}, Z_1^{obs} = z_1, \dots, X_{T-1}^{obs}, Z_{T-1}^{obs} = z_{T-1}). \end{aligned} \quad (3)$$

Because the RHS of (3) involves only distributions of observed variables, relation (3), also called g-computation, establishes estimability of the causal effects in this setting. So, if likelihood models are posited for the distributions involved in the above formula using the original full covariate data, the MLE will generally be more efficient, compared to other methods, when the models are correctly specified.

However, longitudinal treatments increased the need to reduce the dimension of variables to be modeled in the outcome distribution (3). Under assumptions (1, 2, 3), it is a direct consequence of the result for single time treatments that

$$\{Y_i(z) : \text{all } z\} \perp\!\!\!\perp 1(Z_{i,t}^{obs} = k) | e_{i,t,k} \quad (4)$$

where $1(\cdot)$ is the indicator function. This property then suggests that the propensity scores for longitudinal treatments can be used in analogous ways to those in single time treatments. This is a different use of propensity scores for longitudinal treatments than that in weighting methods (e.g. Robins et al. 2000), with or without further adjustments.

Our two goals are the following:

1. Develop an analog to the regression relation (3) that reduces the dimension of the covariates X_t^{obs} to the longitudinal propensity scores. This will be useful because when the dimension of the covariates is reduced, models for the regressions on the propensity scores can be built and checked more easily.
2. Find the admissible estimators if we only keep propensity scores in lieu of the covariates.

At the end of Section 4 we discuss why this admissibility result is a useful guide to what estimation methods are appropriate when we reduce the covariates X_t^{obs} to the propensity scores. Next we address these goals.

4. Regression on longitudinal propensity scores

Consider the setting where the propensity scores are known, and only those scores are kept as summaries of the covariates X_t^{obs} . First, we wish to find how the target quantities $pr(Y(z))$ can be expressed as functions of parameters of likelihood models given the summary information in the propensity scores (and not given X_t^{obs}). Towards that objective, result (4) is not directly useful because the conditioning does not include all past observed treatments assignments.

To assist our objective, we have obtained the following result:

Result 1

If the sequential ignorability assumption given the full covariate history $H_{i,t}$ is true (i.e., Assumption 2), then sequential ignorability remains true when the history of the full covariates is replaced by the history of propensity scores. Specifically, for any particular longitudinal treatment $z = (z_1, \dots, z_T)$, and under assumptions (1, 2, 3), for $t = 1, \dots, T$

$$Y_i(z) \perp\!\!\!\perp 1(Z_{i,t}^{obs} = z_t) | e_{i,1,z_1}, Z_{i,1}^{obs} = z_1, \dots, e_{i,t-1,z_{t-1}}, Z_{i,t-1}^{obs} = z_{t-1}, e_{i,t,z_t}. \quad (5)$$

The result is proven in Appendix A.1 The importance of (5) is its resemblance to Assumption 2. Specifically, the key in deriving the g-computation (3) was Assumption 2. Since the only difference between Assumption 2 and (5) is that the longitudinal covariates in $H_{i,t}$ are replaced by the longitudinal propensity scores, it is easy to show that we make this replacement also in expression (3). Specifically we have:

Result 2

For any longitudinal treatment $z = (z_1, \dots, z_T)$, and under Assumptions (1, 2, 3),

$$\begin{aligned} \text{pr}\{Y(z)\} = \sum_{e_{1,z_1}, \dots, e_{T,z_T}} \text{pr}(Y^{obs} | e_{1,z_1}, Z_1^{obs} = z_1, \dots, e_{T,z_T}, Z_T^{obs} = z_T) \times \\ \text{pr}(e_{1,z_1}) \dots \text{pr}(e_{T,z_T} | e_{1,z_1}, Z_1^{obs} = z_1, \dots, e_{T-1,z_{T-1}}, Z_{T-1}^{obs} = z_{T-1}) \end{aligned} \quad (6)$$

Thus, given models for the RHS of (6),

$$\text{pr}(Y^{obs} | e_{1,z_1}, Z_1^{obs} = z_1, \dots, e_{T,z_T}, Z_T^{obs} = z_T) \text{ and}$$

$$\text{pr}(e_{1,z_1}) \dots \text{pr}(e_{T,z_T} | e_{1,z_1}, Z_1^{obs} = z_1, \dots, e_{T-1,z_{T-1}}, Z_{T-1}^{obs} = z_{T-1}) \quad (7)$$

we can estimate the target quantities $\text{pr}\{Y(z)\}$. Models for (7) can be as flexible as needed, for example splines on functions of the propensity scores (e.g., logits, as in Little and An, 2004) or dummy indicators defined based on propensity scores (e.g., Kang and Schafer, 2007).

The above results are practically important because of the following facts.

Consider first the case where the models for (7) are correct. Then, the Complete Class Theorem of decision theory (Chernoff and Moses, 1960; Ferguson, 1967) proves that, for *any* loss function, any admissible estimator of $\text{pr}\{Y(z)\}$ (a frequentist criterion) must be a Bayes estimator for some prior for the parameters of the likelihood models in (7). In other words, an estimator that is not Bayes for any such prior will be *inadmissible* (see Appendix 4).

Now, Bayes estimators from a likelihood on (7) will depend on the propensity scores only through the sufficient statistics of (7); they will not depend in any other way on the actual values of e . So, even in this longitudinal framework, estimators that use the propensity scores in ways other than through the sufficient statistics of (7), e.g., estimators using weights, cannot be Bayes for any prior and so are inadmissible.

The general reason for this is not only theoretical but anchored on the following practical observation. The HT uses the actual values of the propensity scores, and not through sufficient statistics. Consider now two groups of people with two different propensity score values: the HT would treat the groups differently, no matter the outcomes. Yet, using a flexible regression method we can first assess the evidence that there is a true underlying difference of the outcomes between the two groups: the less evidence that the two groups have different outcome distributions, the more similarly we can treat these groups (the pairing of the groups' propensity scores becomes a sufficient statistic for their two values). This is the main gain in efficiency that the HT ignores.

Inadmissibility has practical importance (Cox and Hinkley, 2000, p. 432–433). Suppose we wish to evaluate estimators by their frequentist mean squared error (MSE as the loss function). Consider an estimator that uses the propensity scores as weights. Then, by the Complete Class Theorem, there exists at least one Bayes estimator that has smaller *frequentist* MSE than the weighting estimator *for all* true values of the parameters. Moreover, because all Bayes estimators have the same large-sample distribution as the maximum likelihood estimator (MLE), we know that the MLE based on the above models will have a large-sample MSE as good as that of the weighting estimator. In terms of large-sample properties, because the weighting estimator is consistent, its inadmissibility means it has suboptimal large sample variance compared to the MLE of the above models.

The above argument assumes models for (7) are correct. If they are incorrect, then our estimators are biased, increasing the MSE, whereas the inverse weighted estimators will still be consistent if the propensity score models are correctly specified. If there is fear of misspecification, though, first, our methods do not need to be competitors to the inverse weighted estimators, but can actually serve as an augmentation that increases efficiency, as is discussed in Sec. 6.. Second, for such a role as augmentation, our use of longitudinal propensity scores in the g-computation, can be more easily checked and modified to be closer to the corresponding true regression models than the g-computation using the full covariates. This distance matters for the efficiency of an augmented estimator (Appendix A. 3). The ease of model diagnostics for our method will be discussed more in Sections 5.3.

5. Data analysis

5.1 Data

The diabetes data are presented in Segal et al (2007). Here we use our new method to perform a different analysis of the data with model checks and comparisons with other methods. We have data (from prescription drug files and from medical claims files) on utilization of medications, inpatient and outpatient services for 131,714 patients with diabetes filling prescriptions from 6/2005–12/2005 (Table 1). We had access to laboratory

data only for a random subset of the population (10%) due to the costs associated with collecting these data. In the analysis presented here, we did not use the laboratory data. However, we repeated the same analysis within the subset of patients whose laboratory data were available and the results obtained were qualitatively and quantitatively similar. The treatments available were “Insulin”, “Exenatide” (byetta, FDA appr. 4/2005), “Both” or “Other”. The treatment labeled “Both” is a combination of insulin and exenatide; “other” is a combination of oral medications. We had three baseline covariates and 18 time varying covariates. We considered three decision periods of two months each covering the first six months after approval and usage of this new drug.

Our goal is to predict and compare patient outcomes if all patients had been assigned to the “Insulin-Insulin-Insulin” (abv. In^3), “Exenatide-Exenatide-Exenatide” (abv. Ex^3) and “Other-Other-Other” (abv. O^3) longitudinal treatments, while accounting for patient and physician factors affecting both initiation and continued use of the medication (Table 3). Note that these comparisons are not the same as the comparisons between the observed distribution of the “ In^3 ” group versus its observed control (everyone other than “ In^3 ”), or the comparison of the observed distribution of “ Ex^3 ” to its observed control, or even the comparison between these two comparisons, as these methods are known to be biased when treatment-assignment confounding is present. For the purpose of demonstrating our methods, we chose to explore two outcomes that impact on drug coverage decisions: hospitalization rates (Hospadm) and total health care charges (CHARGE).

Notice the challenge posed by the fact that almost all patients received either the “Other” or “Insulin” treatments (Table 1). Only 1,444 (1%) and 288 (0.2%) patients were exposed to the new treatments “Exenatide” and “Both” respectively at any point during the study period. Out of the 131,714 patients, 79 (0.06%) were on the “ Ex^3 ” treatment over the 3 decision periods, 7,674 (6%) were on the “ In^3 ” longitudinal treatment and 115,470 (88%) were on the “ O^3 ” longitudinal treatment.

5.2 Description of the data analysis steps

- 0) For each time point $t = 1, \dots, T$ and for all treatments $k = 1, \dots, K$ we estimated propensity score models $\text{pr}(Z_t^{obs}=k|H_t)=e_{t,k}(H_t;\psi_t)$ as a function of prior histories H_t , with ψ_t being the parameters of the propensity score models. Here, we settled for multinomial generalized logit models without interaction terms, as such terms were not needed for balance. Then, for each time period and for each treatment $z = (z_1, \dots, z_T)$ for which we wanted to estimate $E(Y_i(z))$, we divided patients into two groups using the indicators $\mathbf{1}(Z_t^{obs}=z_t)$ and $\mathbf{1}(Z_t^{obs} \neq z_t)$. We then removed patients in the non-overlapping regions of the propensity scores; this reduces extrapolation (e.g., Rosenbaum and Rubin, 1984; Yanovitzky, Zanutto and Hornik, 2005; Rubin and Thomas, 1996; Imbens, 2000), although it has its own problem, discussed in Sec. 6. We repeat this procedure for each time point and treatment. Subsequently we only keep patients who are in all overlapping regions. The number of patients removed for being outside the overlapping regions was small (between 3% to 11% of the total number in each treatment and time period stratum, and about 5% of the overall sample).

- 1) For each treatment $z = (z_1, \dots, z_T)$ for which we wanted to estimate $E(Y_i(z))$, we conducted the following: We divided all patients selected in step 0) into two groups using the indicators $\mathbf{1}(Z_t^{obs} = z_t)$ and $\mathbf{1}(Z_t^{obs} \neq z_t)$. We extracted $e_{t,z_t}(H_t; \hat{\psi}_t)$ and $1 - e_{t,z_t}(H_t; \hat{\psi}_t)$ for each patient i and time t ;
- 2) We then stratified patients according to their quintile, say e_{it}^{str} , of the propensity score distribution of that time, and checked that within each stratum the distributions of the covariates between the treatment groups $\mathbf{1}(Z_t^{obs} = z_t)$ and $\mathbf{1}(Z_t^{obs} \neq z_t)$ were similar. The stratification and the computation of the F-statistics to check detailed balance are performed as usual with single time point and binary treatment studies.
- 3) Here, we fitted models for (7) based on the propensity score quintiles e_{it}^{str} . Alternatively, one can fit more flexible models based on functions of the continuous propensity score values e_{it} as discussed in Sec. 4. Here we limited our analysis to the quintiles as they provided satisfactory balance for most of the covariates. Specifically, we fitted a proportional odds logistic regression model for the transition probabilities $\text{pr}(e_1^{str}; \beta_1)$ and $\text{pr}(e_t^{str} | Z_1^{obs}, e_1^{str}, \dots, Z_{t-1}^{obs}, e_{t-1}^{str}; \beta_t)$, with $\beta_t, 1 \leq t \leq T$ being the parameters of the transition models; here Z_t^{obs} is a binary random variable equal to $\mathbf{1}(Z_t^{obs} = z_t)$.
- 4) Finally, we fitted a standard linear regression for $E(Y^{obs} | Z_1^{obs}, e_1^{str}, \dots, Z_T^{obs}, e_T^{str}; \xi)$ (indexed by parameters ξ) by maximum likelihood. The average potential outcome $E(Y(z))$ is then estimated based on (6) using

$$\hat{E}(Y(z_1, \dots, z_T)) = \sum_{e_1^{str}, \dots, e_T^{str}} E(Y^{obs} | 1, e_1^{str}, \dots, 1, e_T^{str}; \hat{\xi}) \times \text{pr}(e_1^{str}; \hat{\beta}_1) \times \prod_{t=2}^T \text{pr}(e_t^{str} | 1, e_1^{str}, \dots, 1, e_{t-1}^{str}; \hat{\beta}_t) \quad (8)$$

- 5) We repeated steps 1) to 4) for each of the remaining longitudinal treatment combinations of interest $z = (z_1, \dots, z_T)$.

For each treatment and time period, nearly all the explanatory covariates were well-balanced within propensity score quintile (see Table 2 as an example). The uncertainty estimates (standard errors and confidence intervals) of the LHS of (8) were obtained using the bootstrap method (Table 3 and Figure 1). Specifically, we selected with repetition a bootstrap sample from the population of patients selected in step 0 (i.e. in all overlapping regions) and repeated steps 1) to 5). To reduce computational burden here we treated the propensity scores estimated in step 0) as fixed and did not re-estimate them in the bootstrap sample. With one time point this may not understate the uncertainty because it can be better to use an estimated propensity score than a true one (Rubin and Thomas 1996). For more than one time points, the role of the estimated propensity score is more complex so whether or not it can be treated as fixed is a topic of further work. We used the same bootstrap methodology for the inverse propensity score weighted estimator (Figure 1). As mentioned

by a reviewer, the bootstrap variance estimate may not be very accurate, so it will also be of interest to develop estimates based on the theory of the models.

Alternatively, assuming that the propensity score quintiles e_{it}^{str} are given, we could obtain the asymptotic variance of our estimator using the delta method and a numerical approximation of the gradient of the RHS of (8) as a function of $\xi, \beta_1, \dots, \beta_t$ at $\hat{\xi}, \hat{\beta}_1, \dots, \hat{\beta}_t$. Most standard software packages provide the asymptotic covariance matrices of the estimated vectors of parameters $\hat{\xi}, \hat{\beta}_1, \dots, \hat{\beta}_t$ of the various models used in (8).

We estimated that the total health care charges per month that would have occurred (Table 3 and Figure 1) if all patients had been continually on “Exenatide” compared to if the same patients had been on “Insulin” were higher, with a mean monthly difference of \$874 [95% confidence interval (C.I.) \$-232 to \$1,596]. The odds of hospitalization were also higher with a relative odds of 1.17 [95% C.I. 0.17 to 2.47]. These differences were not statistically significant. We also present the results obtained under a different set of models in Table 3 to give an example of model checking and to explore the influence of the models on the results. In Figure 1, we present the results from the comparison between our proposed estimator and the Horvitz-Thompson inverse propensity score weighting estimator.

5.3 Model Checking

A key feature of using regression on the longitudinal propensity score is that we can more easily check the relative validity of the models used in (7). Using the longitudinal propensity score categories e_{it}^{str} in (8) instead of the original multivariate covariates makes the model checking steps straightforward.

We compared our estimates $E(Y^{obs}|1, e_1^{str}, \dots, 1, e_T^{str}; \hat{\xi})$ and $\text{pr}(e_t^{str}|1, e_1^{str}, \dots, 1, e_{t-1}^{str}; \hat{\beta}_t)$ to the observed values under 2 different sets of models. The first set of models included various two and three way interaction terms selected to optimize the exact Akaike Information Criterion (AIC) based on stepwise selection. The second set of models were similar to the first sets except that the outcome models for $E(Y^{obs}|1, e_1^{str}, \dots, 1, e_T^{str})$ were fitted using only patients with $(Z_1^{obs}, Z_2^{obs}, Z_3^{obs}) = (1, 1, 1)$; essentially saturating the “In³” and “Ex³” outcome models. The first set of models can be useful in situations where we have a relatively small number of patients with $(Z_1^{obs}, Z_2^{obs}, Z_3^{obs}) = (1, 1, 1)$ or when we want to borrow strength across longitudinal treatments. The second set of models can be useful when we have a large number of patients with $(Z_1^{obs}, Z_2^{obs}, Z_3^{obs}) = (1, 1, 1)$. The results of the second set were better and are shown in Figure 2. There, though, the observed data are very sparse to justify a saturated model for the estimation of $E(Y^{obs}|1, e_1^{str}, \dots, 1, e_T^{str}; \hat{\xi})$ for patients receiving the “Ex³” or the “In³” longitudinal treatments (“Exenatide Charge $E(Y)$ ” and “Insulin Charge $E(Y)$ ” panels in Figure 2), even though the apparent fit for these patients by definition works better. Thus our models allow direct insights into finding what parts of data carry more versus less information, and thereby offer guidance to better modelling in practice.

6. Discussion

If we trust the likelihood models (7) to be correct, we discussed in Section 4 why, in order for estimators to have good overall accuracy such as mean squared error, the propensity score should be used as regressor.

In the case where the likelihood models (7) are allowed to be incorrect, we have argued and demonstrated through the application in Sec. 5 how the reduced dimensionality of these models makes them relatively easy to check and improve. For single time treatments, this advantage had been demonstrated, e.g., by Dehejia and Wahba (1999) who show that using propensity scores in matching - the saturated form of regression - can provide dramatic benefits for model checking and overall accuracy when compared to using the full covariates.

As a consequence of these checks, the models using longitudinal propensity scores as regressors can be modified to be closer to correct. Thus, the regression estimators we developed can, by continuity with respect to the case when the models are correct (Sec. 4), have overall accuracy closer to the latter ideal case, and thus often better than simple weighted estimators.

Viewed more generally, our proposal to use propensity scores as regressors can be combined, and does not compete, with methods that use both weighting and regression such as from the perspective of semiparametric theory. In particular, when the regression is allowed to be incorrect, theoretical arguments and simulation evidence (e.g., Lunceford and Davidian, 2004) suggest that, in order for an estimator to have good accuracy, it should still use, at least in part, the possibly incorrect model of the outcome on treatment and covariates. In this case, the estimator's precision (not bias) generally deteriorates with increasing distance of the incorrect outcome model from the true one (see Appendix A.3). Since this distance is expected to be large when trying to model too many covariates through an incorrect model, it becomes important that, even in this case, one makes use (e.g., in the outcome component of an augmented weighted estimator) of an outcome model given the propensity score to reduce the covariate dimension per time point, as opposed to an outcome model given all covariates.

This point is demonstrated in Appendix A.3, where we explore in the simple case of the first time (i.e., binary) treatment how the variance of augmented weighted estimators (Rotnitzky, Robins and Scharfstein, 1998; Tsiatis, 2006) is related to the choice of the outcome regression model component when such model is incorrect: with full covariates, or with propensity score only (Figure 1 of Appendix, augmented estimators are used only for this discussion). In this example, for the effects on charge, the variances (evaluated using the bootstrap) are comparable; but for the effects on hospitalization the augmented weighted estimator that uses the propensity score as covariate in the outcome regression is more efficient than the augmented weighted estimator that uses a full covariate outcome regression. Developing theory for augmentations with propensity score regressions in multiple times is a topic of future work.

Our method also has limitations. Specifically, the physician may be interested to know if changing the treatment according to the evolving covariates (a dynamic regime, e.g., Murphy, Van Der Laan, and Robins, 2001) can result in even better outcomes than a fixed, sustained treatment. A set of covariates that we only keep through their reduction to a propensity score is a set that we cannot also use to inform a dynamic regime. This implies that if we want to use a subset of covariates $D_{i,t}$ to inform a dynamic regime, we must use them eventually also as regressors. In such case, the way in which our method gets modified is that we can replace the propensity score $e_{i,t}$ in (5) by $(e_{i,t}$ and $D_{i,t})$. Then, we can get dynamic regimes that depend on $D_{i,t}$ and summarize over $e_{i,t}$, i.e., the remaining covariates.

Also, as with the general use of the g-computation formula, one has to be aware of the possibility of a fact known as the g-null paradox (Robins, Greenland and Hu, 1999). This is the fact that a class of misspecified models for the components in the g-computation formula may not contain a distribution that satisfies the null hypothesis of no causal effects. The extent to which this is a concern is dependent on the extent to which the model is incorrect. Thus, for a method that makes it easier to check and fix the models such as with using a reduced dimension of the covariates with the propensity score, the degree to which the g-null paradox can affect the type I error gets reduced. In addition, the g-null paradox is not a concern when our method is used in combination with a weighting method to preserve consistency but increase precision, as with the augmented estimators discussed in Appendix A.3.

A related issue is that the proportional odds models fitted separately for each treatment (step 4, Sec 5.2) may be theoretically incompatible. Such misspecification could be reduced if we model only the arms of interest. This would not be much of a practical concern when model checking as done in Sec. 5.3 is satisfactory. Nevertheless misspecification concerns with our approach still increase with the number of measurement times, though less so that a full covariate modeling approach. At the expense of possible loss of efficiency, such concerns about the modelling are avoided with a weighted estimator.

It is also important to recognize that the removal of patients who do not overlap in terms of the propensity scores in the longitudinal setting is not as innocuous as in the single time point, because here it is based on pretreatment factors as well as factors possibly affected by treatment. Thus, the propensity score diagnostics are still useful for diagnosing the nonoverlapping regions, but unlike in the single time point, there is not yet a good solution to the possible bias. It may be reassuring though that, when we conducted the analyses described in Sec. 5 without removing the non-overlapping regions (5% of the overall sample) the results (Appendix A.2) were relatively similar to those reported in Table 3.

Finally, one has to also be aware of the problems that can arise in relation to mismatches in timing of the measurements. Specifically, the researcher must trust that the measurements to be “labelled” as occurring at time t are in fact those that were taken into consideration for deciding the treatment at time t . This design concern is important for the validity of any method of estimation of longitudinal treatment effects.

In conclusion, whether or not we expect an outcome regression model to be precisely correct, it is important to have the flexibility to estimate the treatment effects by using the propensity score as a regressor, especially in longitudinal studies where the dimension of covariates increases considerably.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the Editor, the Associate Editor and two reviewers for very helpful comments, and NIH-NIDA R01DA023879-01 for partial support.

References

1. Chernoff, H.; Moses, LE. Elementary decision theory. 2nd ed. New York: Wiley; 1960.
2. Cox, DR.; Hinkley, DV. Theoretical statistics. CRC Press; 2000.
3. Dehejia R, Wahba S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J. Am. Statist. Ass. 1999; 94:1053–1062.
4. TS Ferguson, TS. Mathematical statistics: a decision theoretic approach. New York: London: Academic Press; 1967.
5. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J. Am. Statist. Ass. 1952; 47:663–685.
6. Imbens G. The role of the propensity score in estimating dose-response functions. Biometrika. 2000; 87(3):706–710.
7. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science. 2007; 22:523–580. (with discussion).
8. Little RJ, An H. Robust likelihood based analysis of multivariate data with missing values. Statistica Sinica. 2004; 14:949–968.
9. 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. [PubMed: 15351954]
10. Murphy SA, Van Der Laan MJ, Robins JM. Marginal mean models for dynamic regimes. Journal of the American Statistical Association. 2001; 96:1410–1423. [PubMed: 20019887]
11. Neyman J. On the application of probability theory to agricultural experiments. Essay on Principles. Section 9, translated. Statistical Science. 1923; 5(4):465–480. (with discussion), 1990.
12. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. Journal of Chronic Diseases. 1987; 40:1395–1615.
13. Robins JM, Greenland S, Hu F-C. Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. Journal of the American Statistical Association. 1999; 94:687712. (with discussion).
14. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11:550–560. [PubMed: 10955408]
15. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.
16. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J. Am. Statist. Ass. 1984; 79:516–524.
17. Rotnitzky A, Robins JM, Scharfstein DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. J. Am. Statist. Ass. 1998; 93:1321–1339.

18. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974; 66:688–701.
19. Rubin DB. Bayesian inference for causal effects: The Role of Randomization. *Annals of Statistics*. 1978; 6:34–58.
20. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics*. 1996; 52:249–264. [PubMed: 8934595]
21. Segal JB, Griswold M, Achy-Brou AC, et al. Using propensity scores subclassification to estimate effects of longitudinal treatments: an example using a new diabetes medication. *Medical Care*. 2007; 45:S149–S157. [PubMed: 17909374]
22. Tsiatis, AA. Semiparametric theory and missing data (page 336). New York: Springer-Verlag; 2006.
23. Yanovitzky I, Zanutto E, Hornik R. Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning*. 2005; 28:209–220.

(a) Relative risks of hospitalization

(b) Difference in total charges

(c) Relative risks of hospitalization

(d) Difference in total charges

Figure 1.

Comparing estimators of the relative risk of hospitalization and difference in average monthly total health care charge in a longitudinal causal framework (Jul–Nov05). RLP MC = regression on longitudinal propensity score with AIC based model checking, RLP SAT = regression on longitudinal propensity score with AIC based model checking with

$E(Y^{obs}|1, e_1^{str}, \dots, 1, e_T^{str})$ fitted using only patients with $(Z_1, Z_2, Z_3) = (1, 1, 1)$ and IPW = inverse propensity score weighting. OOO/III = Other-Other-Other versus Insulin-Insulin-Insulin, EEE/III = Exenatide-Exenatide-Exenatide versus Insulin-Insulin-Insulin. The circle represent the point estimates and the bar represents the 95% confidence intervals. The vertical axis is both for point estimates on the left and right of the vertical line in the middle (demarcation line).

Figure 2.

Checking the prediction accuracy of the model selected from the second set described in Sec. 5.3. Here, the models for $E(Y^{obs} | 1, e_1^{str}, \dots, 1, e_T^{str})$ were fitted using only patients with $(Z_1, Z_2, Z_3) = (1, 1, 1)$. Here pe_k observed is $\text{pr}(e_k^{str} | 1, e_1^{str}, \dots, 1, e_{k-1}^{str})$ while pe_k estimate is $\text{pr}(e_k^{str} | 1, e_1^{str}, \dots, 1, e_{k-1}^{str}; \hat{\beta}_k)$. Here, n is the corresponding number of patients and $e1$ is e_1^{str} . We can easily check the models by verifying that most circles are close to the line with a special emphasis on large circles.

Table 1

Drug regimens by treatment period.

July-Aug 05		Sep-Oct 05		Nov-Dec 05		Total	
		Other	Insulin	Exenatide	Both		
Other	Other	115470	1424	677	5	117576	
	Insulin	800	1425	19	11	2255	
	Exenatide	102	6	356	5	469	
	Both	1	0	1	8	10	120310
Insulin	Other	1082	885	26	15	2008	
	Insulin	1268	7674	39	103	9084	
	Exenatide	9	4	29	9	51	
	Both	4	14	12	57	87	11230
Exenatide	Other	21	3	4	0	28	
	Insulin	1	3	0	0	4	
	Exenatide	19	1	79	2	101	
	Both	0	1	0	5	6	139
Both	Other	2	1	0	0	3	
	Insulin	0	2	0	0	2	
	Exenatide	1	0	4	2	7	
	Both	2	3	4	14	23	35
Total		118782	11446	1250	236		131714

Table 2

Balance of covariates between patients filling “Exenatide” and everybody else before and after subclassification into propensity score quintiles (May–June 2005). This table is included here as an example to illustrate the balance among subclasses. The table illustrates that we achieve detailed balance for all but the first 5 covariates. The reduction in F-statistics post stratification is dramatic.

	Before subclassification		After subclassification	
	Fstat	Pvalue	Fstat	Pvalue
endocrine visit jun05	361.10	0.00	256.26	0.00
CHARGE may05	13.45	0.00	9.84	0.00
endocrine visit may05	15.52	0.00	7.23	0.01
anyop visit jun05	36.69	0.00	6.65	0.01
hosplos may05	5.72	0.02	6.36	0.01
rxcopay jun05	69.57	0.00	3.64	0.06
CHARGE jun05	10.60	0.00	2.63	0.10
anyop visit may05	8.44	0.00	1.27	0.26
nrx jun05	53.00	0.00	0.50	0.48
phtc	3.97	0.05	0.43	0.51
hospadm may05	0.00	0.96	0.30	0.59
thiazolidione may05	10.77	0.00	0.19	0.66
rxcopay may05	13.43	0.00	0.12	0.73
thiazolidione jun05	8.68	0.00	0.10	0.76
npts	0.26	0.61	0.08	0.78
obesity jun05	1.15	0.28	0.06	0.80
gender	8.11	0.00	0.06	0.81
fampract visit jun05	5.22	0.02	0.06	0.81
intmed visit may05	0.03	0.85	0.06	0.81
comorb jun05	5.35	0.02	0.04	0.84
ervisit may05	0.28	0.59	0.03	0.87
obesity may05	0.69	0.41	0.02	0.88
comorb may05	4.15	0.04	0.02	0.89
ervisit jun05	0.49	0.48	0.02	0.90
nrx may05	16.50	0.00	0.01	0.92
sulfonyureas jun05	8.44	0.00	0.01	0.92
intmed visit jun05	7.96	0.01	0.01	0.92
fampract visit may05	0.34	0.56	0.01	0.93
side effects may05	0.16	0.69	0.00	0.95
o med may05	3.93	0.05	0.00	0.95
o med jun05	6.04	0.01	0.00	0.96
sulfonyureas may05	3.06	0.08	0.00	0.97
metformin jun05	8.39	0.00	0.00	0.98
metformin may05	1.51	0.22	0.00	0.98
age	0.06	0.81	0.00	0.99

Table 3

Average monthly hospitalization frequency, relative risk of hospitalization, average monthly total health care charges and difference in average monthly total health care charge in a longitudinal causal framework (Jul–Nov05) using regression on the longitudinal propensity score with AIC based model checking.

(a) Treatment effects for “Hospadm”.

Medication in each time period	Hospadm	95% CI	RR	95% CI
Insulin-Insulin-Insulin	0.0083	(0.0070, 0.0101)	1	reference
Other-Other-Other	0.0101	(0.0098, 0.0104)	1.213	(1.012, 1.488)
Exenatide-Exenatide-Exenatide	0.0098	(0.0015, 0.0204)	1.171	(0.174, 2.474)

(b) Treatment effects for “Hospadm”. Here, the models for $E(Y^{obs}|1, e_1^{str}, \dots, 1, e_T^{str})$ were fitted using only patients with $(Z_1, Z_2, Z_3) = (1, 1, 1)$.

Medication in each time period	Hospadm	95% CI	RR	95% CI
Insulin-Insulin-Insulin	0.0082	(0.0073, 0.0095)	1	reference
Other-Other-Other	0.0101	(0.0099, 0.0104)	1.240	(1.066, 1.387)
Exenatide-Exenatide-Exenatide	0.0024	(0.00001, 0.0064)	0.293	(0.002, 0.690)

(c) Treatment effects for “CHARGE”.

Medication in each time period	CHARGE (\$)	95% CI	Diff.	95% CI
Insulin-Insulin-Insulin	840	(785, 901)	0	reference
Other-Other-Other	976	(961, 992)	135	(74, 184)
Exenatide-Exenatide-Exenatide	1715	(640, 2427)	874	(−232, 1596)

(d) Treatment effects for “CHARGE”. Here, the models for $E(Y^{obs}|1, e_1^{str}, \dots, 1, e_T^{str})$ were fitted using only patients with $(Z_1, Z_2, Z_3) = (1, 1, 1)$.

Medication in each time period	CHARGE (\$)	95% CI	Diff.	95% CI
Insulin-Insulin-Insulin	845	(796, 905)	0	reference
Other-Other-Other	974	(962, 994)	129	(80, 189)
Exenatide-Exenatide-Exenatide	1276	(527, 2511)	432	(−337, 1654)