*Article*

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(3) 709–727

DOI: 10.1177/0962280219888745
journals.sagepub.com/home/smm
SAGE

# Propensity score-based methods for causal inference in observational studies with non-binary treatments

**Shandong Zhao[1], David A van Dyk[2] and Kosuke Imai[3]**

**Abstract**

Propensity score methods are a part of the standard toolkit for applied researchers who wish to ascertain causal effects from observational data. While they were originally developed for binary treatments, several researchers have proposed generalizations of the propensity score methodology for non-binary treatment regimes. Such extensions have widened the applicability of propensity score methods and are indeed becoming increasingly popular themselves. In this article, we closely examine two methods that generalize propensity scores in this direction, namely, the propensity function (PF), and the generalized propensity score (GPS), along with two extensions of the GPS that aim to improve its robustness. We compare the assumptions, theoretical properties, and empirical performance of these methods. On a theoretical level, the GPS and its extensions are advantageous in that they are designed to estimate the full dose response function rather than the average treatment effect that is estimated with the PF. We compare GPS with a new PF method, both of which estimate the dose response function. We illustrate our findings and proposals through simulation studies, including one based on an empirical study about the effect of smoking on healthcare costs. While our proposed PF-based estimator preforms well, we generally advise caution in that all available methods can be biased by model misspecification and extrapolation.

**Keywords**

Covariate adjustment, generalized propensity score, model diagnostics, propensity function, smooth coefficient model, stabilized weights, nonparametric models

## 1 Introduction

One of the most common strategies used in numerous scientific disciplines to make causal inferences in observational studies is to adjust for observed confounding variables. Researchers find that results based on regression adjustments, however, can be sensitive to model specification, especially when applied to data where the treatment and control groups differ substantially in terms of their pre-treatment covariates. The propensity score methods of Rosenbaum and Rubin,[1] hereafter RR, aim to address this fundamental problem by reducing the covariate imbalance between the two groups. RR showed that under the assumption of no unmeasured confounding, adjusting for the propensity score, rather than potentially high-dimensional covariates, is sufficient for unbiased estimation of causal effects and this can be accomplished using simple nonparametric methods such as matching and subclassification.

[1]Department of Statistics, University of California, Irvine, CA, USA
[2]Department of Mathematics, Imperial College London, UK
[3]Department of Government and Department of Statistics, Harvard University, Cambridge, MA, USA

**Corresponding author:**
Kosuke Imai, Department of Government and Department of Statistics, Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA.
Email: Imai@Harvard.Edu

Despite their popularity, one limitation of the original propensity score methods is that they are only applicable to studies with a treatment and control group, that is, studies with a binary treatment. Some years ago, several researchers proposed generalization of the propensity score methodology for non-binary treatment regimes.[2–5] Such extensions have widened the applicability of propensity score methods and are indeed becoming increasingly popular themselves.

These methods, however, require users to overcome the challenges of (i) correctly modeling a treatment variable as a function of a possibly large number of pre-treatment covariates and (ii) modeling the response variable. Both represent significant challenges in practice despite recent progress on more robust modeling strategies.[6,7] Standard diagnostics based on the comparison of the covariate distributions between the treatment and control groups are not directly applicable to non-binary treatment regimes and the final inference can be quite sensitive to the choice of response model. This motivated Flores et al.,[8] hereafter FFGN, to propose two extensions to the method of Hirano and Imbens[2] that aim to provide more robust estimation through a more flexible response model.

In this article, we closely examine two propensity score-based methods for causal inference with non-binary treatments, namely, the propensity function (PF) of Imai and van Dyk,[3] hereafter IvD; and the generalized propensity score (GPS) of Hirano and Imbens,[2] hereafter HI, along with the FFGN extensions. We compare the assumptions and theoretical properties of these methodologies when used to estimate a dose–response function (DRF) and examine their empirical performance in practice. Our primary message is cautionary: estimating the full DRF with a continuous treatment in an observational study is challenging. Researchers should be cautious when attempting to do so.

The methods of IvD and HI each formalize a framework for inference, but are based on ideas that appear earlier in the literature.[4,9–11] In this article we focus on the estimation of the DRF in an observational study in which both the covariates and treatment are fixed, as outlined in HI. This is the simplest setting beyond estimating an average treatment effect and is a well-defined testing ground for the clear comparison of available methods. As we shall see, even in this simple setting standard methods may exhibit unacceptable statistical properties.

The methods of IvD and HI are not the only ones that can be used to estimate the effect of non-binary treatments in observational studies. Inverse probability weighting (IPW)[5] (hereafter RHB) for example, is a general method that (like GPS) enjoys wide application beyond the estimation of a DRF, such as estimating the effect of time-varying treatments and with longitudinal data.[12–14] Ertefaie and Stephens[15] provides a comparison of IPW and GPS in this setting. In part due to space constraints, we focus on using GPS and PF to estimate the DRF, and discuss IPW only insofar as it comes into one of the extensions of the GPS proposed by FFGN.

The remainder of this article is organized into five sections. Section 2 reviews the theoretical properties of the original propensity score methodology and its interrelated generalizations. While the GPS is designed to estimate the DRF, the PF estimates the average treatment effect. In section 3, we compare the methods of IvD, HI, and FFGN both theoretically and empirically. We demonstrate that the response model used by HI is less flexible than those typically used with propensity score methods and that the methods proposed by FFGN to address this problem can exhibit undesirable properties. We also show that one of FFGN's methods can be improved by using the stabilized weights of RHB, effectively implementing IPW to estimate the full DRF. In section 4, we compare these methods with a new proposal and show how the method of IvD can also be extended for robust estimation of the full DRF. The efficacy of the proposed methodology is illustrated through simulation studies in section 4 and an empirically based study in section 5. Section 6 offers concluding remarks while online Appendix 1 presents additional simulation results and online Appendix 2 introduces a robust variant of HI's method. Table 1 lists and defines the abbreviations used in the text.

## 2 Methods

Suppose we have a simple random sample of size $n$ with each unit consisting of a $p$-dimensional column vector of pretreatment covariates, $\boldsymbol{X}_i$, the observed univariate treatment, $T_i$, and the outcome variable, $Y_i$. Although IvD's method can be applied to multivariate treatments, here we assume the treatment is univariate to facilitate comparison with the method of HI. We omit the subscript when referring to generic values of $\boldsymbol{X}_i$, $T_i$, and $Y_i$.

We denote the potential outcomes by $\mathcal{Y} = \{Y_i(t), t \in \mathcal{T} \text{ for } i = 1, \ldots, n\}$, where $\mathcal{T}$ is a set of possible treatment values and $Y_i(t)$ is a function that maps a particular treatment level of unit $i$, to its outcome. This setup implies the *stable unit treatment value assumption*[16] that the potential outcome of each unit is not a function of treatment level of other units and that the same version of treatment is applied to all units. In addition, we assume *strong ignorability of treatment assignment*, i.e. $Y(t) \perp\!\!\!\perp T|\boldsymbol{X}$ and $p(T = t|\boldsymbol{X}) > 0$ for all $t \in \mathcal{T}$, which implies no unmeasured confounding (RR).

**Table 1.** List of abbreviations used in text.

| Abbreviation | Meaning |
|---|---|
| DRF | dose response function |
| FFGN | the paper by Flores, Flores-Lagunes, Gonzalez and Neumann[8] |
| GPS | generalized propensity score |
| HI | the paper by Hirano and Imbens[2] |
| IPW | inverse probability weighting |
| IvD | the paper by Imai and van Dyk[3] |
| PF | propensity function |
| RHB | the paper by Robins, Hernán and Brumback[5] |
| RR | the paper by Rosenbaum and Rubin[1] |
| SCM | smooth coefficient model |

## 2.1 The propensity score with a binary treatment

RR considered the case of treatment variables that take on only two values, $\mathcal{T} = \{0, 1\}$, where $T_i = 1$ ($T_i = 0$) implies that unit $i$ receives (does not receive) the treatment and defined the *propensity score* to be the conditional probability of assignment to treatment given the observed covariates, i.e. $e(\boldsymbol{X}) = p(T = 1|\boldsymbol{X})$. In practice, $e(\boldsymbol{X})$ is typically estimated using a parametric treatment assignment model $p_\psi(T = 1|\boldsymbol{X})$ where $\psi$ is a vector of unknown parameters. The appropriateness of the fitted model can be assessed via the celebrated balancing property of $e(\boldsymbol{X})$, namely, that covariates should be independent of the treatment conditional on the propensity score, $\boldsymbol{X} \perp\!\!\!\perp T|e(\boldsymbol{X})$. In particular, the fitted model, $\hat{e}(\boldsymbol{X}) = p\hat{\psi}(T = 1|\boldsymbol{X})$ should not be accepted unless adjusting for $\hat{e}(\boldsymbol{X})$ results in adequate balance.

In order to estimate causal quantities, we must properly adjust for $\hat{e}(\boldsymbol{X})$. RR propose three techniques: matching, subclassification, and covariance adjustment. We focus on subclassification and covariance adjustment because they are more closely related to the generalizations for non-binary treatments. The key advantage of propensity scores when applying these methods, and the inverse weighting method discussed below, is dimension reduction. They only require adjustment for a scalar variable $\hat{e}(\boldsymbol{X})$ rather than for the entire covariate vector.

With subclassification (RR), we adjust for $\hat{e}(\boldsymbol{X})$ by dividing the observations into several subclasses based on $\hat{e}(\boldsymbol{X})$. Individual response models are then fitted within each subclass, adjusting for $\hat{e}(\boldsymbol{X})$ and sometimes $\boldsymbol{X}$ along with $T$. The overall causal effect is then computed as the weighted average of the within-class coefficients of $T$, with weights proportional to the size of subclass. The standard error of the causal effect is typically computed by treating the within-subclass estimates as independent of one another.

With covariance adjustment (RR), we regress the response variable on $\hat{e}(\boldsymbol{X})$ separately for the treatment and control groups. Specifically, we divide the data into the treatment and control groups and fit the regression model, $\mathrm{E}(Y|\boldsymbol{X}, T = t) = \alpha_t + \beta_t \cdot \hat{e}(\boldsymbol{X})$, separately for $t = 0$, 1. The average causal effect is then estimated as

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\alpha}_1 + \hat{\beta}_1 \cdot \hat{e}(\boldsymbol{X}_i) - \hat{\alpha}_0 - \hat{\beta}_0 \cdot \hat{e}(\boldsymbol{X_i})\right) = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0) \cdot \overline{\hat{e}(\boldsymbol{X})} \tag{1}$$

where $\overline{\hat{e}(\boldsymbol{X})}$ is the sample mean of the estimated propensity score.

Subclassification and covariance adjustment (along with methods such as matching and IPW) aim to provide robust flexible adjustment for $\hat{e}(\boldsymbol{X})$ in the response model. As we shall see below, the flexibility of the response model is important especially for non-binary treatment regimes. This is because unlike the treatment assignment model which has an effective diagnostic tool based on the balancing property of propensity scores, the response model lacks such diagnostics.

## 2.2 Propensity score: methods for non-binary treatments

Suppose now that $\mathcal{T}$ is a more general set of treatment values, perhaps categorical or continuous. It is in this setting that that IvD introduced the PF and that HI introduced the GPS. (IvD also allow for multi-variate treatments, which we do not discuss in this paper.) In what follows, we review and compare these generalizations

of propensity score methods. In particular, we consider the following aspects of propensity score adjustment with binary treatments that both IvD and HI generalize:

1. Treatment assignment model: Model the distribution of the treatment assignment given covariates to estimate the propensity score, i.e. $\hat{e}(\boldsymbol{X})$.
2. Diagnostics: Validate $\hat{e}(\boldsymbol{X})$, by checking for covariate balance, i.e. $T \perp\!\!\!\perp \boldsymbol{X} | \hat{e}(\boldsymbol{X})$.
3. Response model: Model the distribution of the response given the treatment, adjusting for $\hat{e}(\boldsymbol{X})$ via matching, subclassification, covariance adjustment, or IPW.
4. Causal quantities of interest: Estimate the causal quantities of interest and their standard error based on the fitted response model.

#### 2.2.1 Treatment assignment model

As in the case of the binary treatment, we begin by modeling the distribution of the observed treatment assignment given the covariates using a parametric model, $p_\psi(T|\boldsymbol{X})$, where $\psi$ is a set of parameters. Common choices of $p_\psi(T|\boldsymbol{X})$ include the Gaussian or multinomial regression models when the treatment variable is continuous or categorical, respectively. HI define the GPS as $R = r(T, \boldsymbol{X}) = p_\psi(T|\boldsymbol{X})$. That is, the GPS is equal to the treatment assignment model density (or mass) function *evaluated* at the observed treatment variable and covariate for a particular individual. This is analogous to the propensity score for the binary treatment, which can be written as $e(\boldsymbol{X}) = r(1, \boldsymbol{X}) = p_\psi(T = 1|\boldsymbol{X})$.

Before HI coined the term GPS, the same quantity was used by RHB in IPW. In particular, RHB considered using weights equal to $1/r(T, \boldsymbol{X})$ and then, noting the instability of these weights, suggested using the stabilized weights $W = W(T, \boldsymbol{X}) = p_{\psi_0}(T)/p_\psi(T|\boldsymbol{X})$ instead. For example, if we use a normal linear model for $p_\psi(T|\boldsymbol{X})$, i.e. $T_i|\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{X}_i^T\beta, \sigma^2)$, we might use a normal model for $p_{\psi_0}(T)$, i.e. $T_i \sim \mathcal{N}(\mu, \tau^2)$. Although RHB first proposed the quantity, we use HI's now standard term, GPS, for $r(T, \boldsymbol{X})$.

IvD summarize $p_\psi(T|\boldsymbol{X})$ in a manner that is qualitatively different. First, they define the PF to be the entire conditional density (or mass) function of the treatment, namely $e_\psi(\cdot|\boldsymbol{X}) = p_\psi(\cdot|\boldsymbol{X})$. This is also analogous to the propensity score for the binary treatment case because $e_\psi(\cdot|\boldsymbol{X})$ is completely determined by $e(\boldsymbol{X}) = p_\psi(T = 1|\boldsymbol{X})$. In order to summarize the PF, IvD introduce the *uniquely parameterized propensity function* assumption which states that for every value of $\boldsymbol{X}$, there exists a unique finite-dimensional parameter, $\boldsymbol{\theta} \in \Theta$, such that $e_\psi(\cdot|\boldsymbol{X})$ depends on $\boldsymbol{X}$ only through $\boldsymbol{\theta}_\psi(\boldsymbol{X})$. In other words, $\boldsymbol{\theta}$ uniquely represents $e\{\cdot|\boldsymbol{\theta}_\psi(\boldsymbol{X})\}$, which we may therefore write as $e(\cdot|\boldsymbol{\theta})$ or simply $\boldsymbol{\theta} = \boldsymbol{\theta}_\psi(\boldsymbol{X_i})$. For example, if we model the treatment, $T_i \sim \mathcal{N}(\boldsymbol{X}_i^\top \boldsymbol{\beta}, \sigma^2)$ with $\psi = (\boldsymbol{\beta}, \sigma^2)$, the scalar $\boldsymbol{\theta_i} = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ uniquely represents $e_\psi(\cdot|\boldsymbol{X_i})$. In practice, $\psi$, $\psi_0$, $\boldsymbol{\theta}_i$, $W_i$, $R_i$, and $r_i$ are estimated from data; we denote their estimates by $\hat{\psi}$, $\hat{\psi}_0$, $\hat{\boldsymbol{\theta}}_i$, $\hat{W}_i$, $\hat{R}_i$, and $\hat{r}_i$, respectively.

#### 2.2.2 Diagnostics

Diagnostics for the treatment assignment model rely on balancing properties of the IPW, PF and GPS. For example, IvD shows that the PF is a balancing score, i.e. $T \perp\!\!\!\perp \boldsymbol{X} | e(\cdot|\boldsymbol{\theta})$. IvD suggest checking balance by regressing each covariate on $T$ and $\hat{\boldsymbol{\theta}}$, e.g. using Gaussian and/or logistic regression and comparing the distribution of the $t$-statistics for each of the resulting regression coefficients of $T$ with the standard normal distribution via a normal quantile plot. Improvement in balance can be assessed by constructing the plot again in the same manner except that $\hat{\boldsymbol{\theta}}$ is left out of each regression. Although not typical used, this diagnostic is equally applicable in the binary treatment case.

HI, on the other hand, show that $\mathbf{1}\{T = t\}$ is independent of $\boldsymbol{X}$ given $r(t, \boldsymbol{X})$, where $\mathbf{1}\{\cdot\}$ is an indicator function and the GPS is evaluated at $t \in \mathcal{T}$. Following the covariate balancing property for the binary propensity score, HI construct a series of binary treatments by coarsening the original treatment $T$ in the form $\{t_j < T \leq t_{j+1}\}$ for some $t_1, t_2, \ldots, t_J$. Covariate balance is then checked for these binary treatment variables by first subclassifying units on $\hat{r}(\tilde{T}_j, \boldsymbol{X})$, where $\tilde{T}_j$ is the median of the treatment variable among units with $\mathbf{1}\{t_j < T \leq t_{j+1}\} = 1$. Then, two-sample $t$-tests are performed within each subclass to compare the mean of each covariate among units with $\mathbf{1}\{t_j < T \leq t_{j+1}\} = 0$ against that among units with $\mathbf{1}\{t_j < T \leq t_{j+1}\} = 1$. Finally the within-subclass differences in means and the variances of these differences are combined to compute a single t-statistic for each covariate. HI suggest repeating this diagnostics for several choices of $\{t_1, \ldots, t_J\}$ that cover the range of observed $T$.

Because the same treatment models are used with both methods the diagnostics for each may be used for the other. In any case, failure to reject the null hypothesis of perfect balance does not imply balance and hence the

diagnostics must be interpreted carefully. In fact, a small (within subclass) sample size, may limit the ability to detect a lack of balance.[17]

#### 2.2.3 Response model

The response models proposed by IvD and HI are quite different, with HI relying more heavily on parametric assumptions. IvD propose two response models. The first is completely analogous to the subclassification technique proposed by RR. The data are subclassified on $\hat{\boldsymbol{\theta}}$ and individual response models are fitted within each subclass, adjusting for $\hat{\boldsymbol{\theta}}$ and typically $\boldsymbol{X}$ along with $T$. The second is a smooth coefficient model (SCM), which allows the intercept and slope to vary smoothly as a function of the PF

$$E(Y|T,\hat{\boldsymbol{\theta}}) = f(\hat{\boldsymbol{\theta}}) + g(\hat{\boldsymbol{\theta}}) \cdot T \tag{2}$$

where $f(\cdot)$ and $g(\cdot)$ are unknown smooth continuous functions. In our numerical illustrations, we fit this model using the R package `mgcv` developed by Simon Wood, in which smooth functions are represented as a weighted sum of known basis functions, and the likelihood is maximized with an added smoothness penalization term. We use penalized cubic regression splines as the basis functions, with dimension equal to five.

In contrast, HI propose to estimate the conditional expectation of the response as a function of the observed treatment, $T$, and the GPS, $\hat{R}$. They recommend using a flexible parametric function of the two arguments and give the following Gaussian quadratic regression model

$$E(Y|T,\hat{R}) = \alpha_0 + \alpha_1 \cdot T + \alpha_2 \cdot T^2 + \alpha_3 \cdot \hat{R} + \alpha_4 \cdot \hat{R}^2 + \alpha_5 \cdot T \cdot \hat{R} \tag{3}$$

This can be viewed as a generalization of RR's covariance adjustment technique, which in the binary treatment case involves regressing $Y$ on $\hat{e}(\boldsymbol{X})$ separately for the treatment and control groups. HI, on the other hand, parametrically estimate the average outcome for all possible treatment levels simultaneously via the quadratic regression on $T$ given in equation (3). Non-parametric response models for the GPS are suggested by FFGN, see section 2.2.5.

#### 2.2.4 Estimating causal quantities

The PF was designed to estimate the average causal effect. Under equation (2) this involves averaging $g(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}})$ across all units. Bootstrap standard errors are computed by resampling the data and refitting both the treatment assignment and response models. With subclassification, computing the estimated average causal effect proceeds exactly as in the binary case. Because a response model is fit conditional on $T$ within each subclass, we can also in principle average these fitted models and estimate the DRF. While we illustrate this possibility in our simulations, we advocate a flexible non-parametric approach in section 4.1.

In contrast, to estimate the DRF, HI computes the average potential outcome on a grid of treatment values. In particular, at treatment level $t$, they compute

$$\hat{E}\{Y(t)\} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_0 + \hat{\alpha}_1 \cdot t + \hat{\alpha}_2 \cdot t^2 + \hat{\alpha}_3 \cdot \hat{r}(t,\boldsymbol{X_i}) + \hat{\alpha}_4 \cdot \hat{r}(t,\boldsymbol{X_i})^2 + \hat{\alpha}_5 \cdot t \cdot \hat{r}(t,\boldsymbol{X_i})) \tag{4}$$

In equation (4) the mean response had all individuals received dose $t$ is estimated by computing a *dose-specific* score: $r$ is evaluated at $t$, not $T$. While not unprecedented,[4,11] this distinguishes the method from the PF, for which there is a single score for each individual. This difference has ramifications. When the inverse GPS is used as a weight, it cannot be stabilized in the standard manner because $p_{\hat{\psi}_0}(t)$ does not vary among individuals so that $p_{\hat{\psi}_0}(t)/r(t,\boldsymbol{X}_i) \propto 1/r(t,\boldsymbol{X}_i)$. Standard errors can be calculated using the bootstrap, taking into account the estimation of both the GPS and model parameters.

In practice, we are often interested in the *relative* DRF, $E\{Y(t) - Y(0)\}$, which compares the average outcome under each treatment level with that under the control, i.e. $t = 0$. Of course, in some studies there is no control *per se* and we revert to $E\{Y(t)\}$. In our simulation studies, we report the relative DRF while in our applied example we report the DRF which is more appropriate in its particular context.

#### 2.2.5 Extensions of the method of HI

Unfortunately, the quadratic regression in equation (3) is not sufficiently flexible for robust estimation of the DRF, see section 3. Bia et al.[18] and FFGN point out that misspecification of equation (3) can result in biased causal quantities and consider similar nonparametric regression methods. In particular, they generalize equation (3) with

$$E(Y|T,\hat{R}) = \beta(T,\hat{R}) \tag{5}$$

where $\beta(T,\hat{R})$ is a flexible nonparametric model; in our numerical studies, we use a SCM.[a] The DRF, $\hat{E}\{Y(t)\}$, and its standard errors are computed as in equation (4), but with

$$\hat{E}\{Y(t)\} = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}[t,\hat{r}(t,\boldsymbol{X}_i)] \tag{6}$$

Because the SCM is a function of the GPS, we refer to this method as SCM(GPS).

FFGN also propose a second method which involves a GPS version of inverse probability weighting. We refer to this method as $\text{IPW}_0$ because it uses naive rather than stabilized weights; recall that stabilization is not possible when using $1/\hat{r}(t,\boldsymbol{X}_i)$ as weights, see section 2.2.4. The $\text{IPW}_0$ estimate of the DRF is

$$\hat{E}\{Y(t)\} = \frac{\sum_{i=1}^{N}\tilde{K}_{h,X}(T_i - t)\cdot Y_i}{\sum_{i=1}^{N}\tilde{K}_{h,X}(T_i - t)} \tag{7}$$

where $\tilde{K}_{h,X}(T_i - t) = K_h(T_i - t)/\hat{r}(t,\boldsymbol{X}_i)$, $K(\cdot)$ is a kernel function with the usual properties, $h$ is a bandwidth satisfying $h \to 0$ and $Nh \to \infty$ as $N \to \infty$, and $K_h(\cdot) = h^{-1}K(\cdot/h)$. This is the local constant regression (Nadaraya–Watson) estimator but now with each individual's kernel weight being divided by its GPS at $t$. To avoid boundary bias and to simplify derivative estimation, the $\text{IPW}_0$ estimates $E\{Y(t)\}$ using a local linear regression of $Y$ on $T$ with a weighted kernel function $\tilde{K}_{h,X}(T_i - t)$, i.e.

$$\hat{E}\{Y(t)\} = \frac{D_0(t)S_2(t) - D_1(t)S_1(t)}{S_0(t)S_2(t) - S_1^2(t)} \tag{8}$$

where $S_j(t) = \sum_{i=1}^{N}\tilde{K}_{h,X}(T_i - t)(T_i - t)^j$ and $D_j(t) = \sum_{i=1}^{N}\tilde{K}_{h,X}(T_i - t)(T_i - t)^j Y_i$. The global bandwidth can be chosen following the procedure of Fan and Gijbels.[19] We use equation (8) as the $\text{IPW}_0$ estimator in our numerical studies.

Unfortunately, $\text{IPW}_0$ can be very unstable, owing to the infinite variance of $1/r(t,\boldsymbol{X}_i)$, at least when the treatment is continuous (RHB). To improve $\text{IPW}_0$ we replace $1/\hat{r}(t,\boldsymbol{X}_i)$ with Robins' stabilized weight $\hat{W}(T,\boldsymbol{X})$. We denote this method $\text{IPW}_{\text{SW}}$, where the subscript indicates its stabilized weights. Notice that these weights are evaluated at $T$ rather than $t$. This is a fundamental difference from the methods laid out by HI. In this regard, $\text{IPW}_{\text{SW}}$ should be viewed as IPW in flexible kernel smoothing regression, namely via equation (7), but with $\tilde{K}_{h,X}(T_i - t) = K_h(T_i - t)\hat{W}(T_i,\boldsymbol{X}_i)$. Table 2 summarizes the specific estimates of the DRF that we compare in sections 3 to 5.

## 3 Comparing the GPS and the PF

In this section, we examine the differences between the methods of IvD, HI, and FFGN using both simulation studies and theoretical comparisons. The key differences lie in how the method summarizes $p(T|\boldsymbol{X})$: the GPS evaluate this density at the observed covariate, whereas the PF uniquely parameterizes it. As we show below, this difference leads to alternative response models and markedly divergent results.

### 3.1 Simulation study I

In our first simulation study, we generate 2000 observations, each of which includes a single continuous covariate, $X$, a continuous univariate treatment, $T$, and a response variable, $Y$. We simulate $X_i \overset{\text{ind}}{\sim} \mathcal{N}(0.5, 0.25)$ and $T_i|X_i \overset{\text{ind}}{\sim} \mathcal{N}(X_i, 0.25)$ and assume that the potential outcome is distributed as $Y_i(t)|T_i, X_i \overset{\text{ind}}{\sim} \mathcal{N}(10X_i, 1)$ for all

**Table 2.** Estimates of the DRF.

| Estimate | Description |
|---|---|
| IvD | Using $\hat{\theta}$ to form $S$ subclasses, fit a linear model within each subclass and average the $S$ fitted models to estimate DRF. |
| HI | Estimate the DRF with equations (3) and (4)[a] |
| SCM(GPS) | Same as HI, but with the quadratic regressions in (3) and (4) replaced by a SCM as in equations (5) and (6) |
| $IPW_0$ | Same as HI, but with the quadratic regressions in equations (3) and (4) replaced by the kernel smoothing regression in equations (7) and (8). |
| $IPW_{SW}$ | Same as $IPW_0$, but with $\tilde{K}_{h,X}(T_i - t) = K_h(T_i - t)\hat{W}(T_i, \boldsymbol{X}_i)$. |
| SCM(PF) | Estimate the DRF with equations (11) and (12), see section 4. |

Note: Methods differ in their response models, but not their treatment models.
[a]In the linear fit of Simulation II, the quadratic models in equations (3) and (4) are replaced by linear models.

$t \in \mathcal{T}$. (Here and elsewhere we parameterize the Normal distribution in terms of its mean and variance.) Under this simulation setup, the true treatment effect is zero and the true DRF is five for all $t$. We deliberately choose this simple setting where any reasonable method should perform well. Fitting a simple linear regression of $Y$ on $T$ yields a statistically significant treatment effect estimate of roughly five. However, adjusting for $X$ in the regression model is sufficient to yield an estimate that is much closer to and is not statistically different from the true effect of zero.

Using the correctly specified treatment assignment model, $T_i|X_i \overset{\text{ind}}{\sim} \mathcal{N}(X_i, 0.25)$, the marginal distribution of the treatment, $T_i \sim \mathcal{N}(0.5, 0.5)$, and the response models given in Table 2, we implement the HI, IvD, SCM(GPS), $IPW_0$, and $IPW_{SW}$ methods. For the purposes of illustration, we do not adjust for $\hat{\theta}$ within each subclass when using IvD's method. Owing to the linear structure of the generative model, doing so would dramatically reduce bias even with a small number of subclasses. Here we illustrate, instead, how bias can be reduced by increasing the number of subclass; we implement IvD with $S = 5$, 10, and 50 subclasses. For the methods other than IvD, we use a grid of 10 equally spaced points between $-0.5$ and 1.5, $t_1, \ldots, t_D$ with $D = 10$, to compute the relative DRF and its derivative. Standard errors are computed using 1000 bootstrap replications.

Figure 1 presents the results. In the first row, we plot the estimated relative DRF while the second row plots the estimated derivative of the DRF. For HI, SCM(GPS), $IPW_0$, and $IPW_{SW}$ the derivative is computed as

$$\frac{1}{2}\left[\frac{\hat{E}\{Y(t_{d+1})\} - \hat{E}\{Y(t_d)\}}{t_{d+1} - t_d} + \frac{\hat{E}\{Y(t_d)\} - \hat{E}\{Y(t_{d-1})\}}{t_d - t_{d-1}}\right] \tag{9}$$

for $d = 2, \ldots D - 1$. For $d = 1$, we simply use the first term in equation (9) and for $d = 10$ we use the second term in equation (9). For IvD, the derivative is the weighted average of the within subclass linear regression coefficient; 95% point-wise confidence intervals are shaded gray.

Figure 1 shows that even in this simple simulation, all methods except $IPW_0$ miss the true relative DRF and its derivative, albeit to differing degrees. The behavior of IvD's estimate improves with more subclasses, a luxury we can afford here because of the large sample size. IvD makes the general recommendation that the within subclass models be adjusted for $\boldsymbol{X}$ or at least for $\hat{\theta}$. Because of the simple structure of this simulation, doing so would result in a correctly specified model even with a single subclass, eliminating bias in the estimated average treatment effect.[b] We do not recommend estimating the DRF by averaging the unadjusted within subclass models, but do so here to facilitate comparisons between the methods. We propose a new estimate of the DRF using the PF in section 4.1.

The performance of HI's method is particularly poor; it differs only slightly from the unadjusted regression. Although SCM(GPS) offers limited improvement, it also introduces a cyclic artifact into the fit. We see this pattern again and discuss it in section 1.3. $IPW_0$, on the other hand, results in an unstable fit that is characterized by large standard errors. The performance of these methods are especially troubling both because the GPS was expressly designed to estimate the DRF and because the current simulation setup is so simple. Given their performance here, it is difficult to expect these methods to succeed in more realistic settings. While $IPW_{SW}$ also misses the mark in Simulation I, it is important to emphasize that RHB did not propose to use IPW to estimate the full DRF. The primary goal of this paper is to explain why the GPS-based methods can fail and to provide a more robust estimate of the DRF.

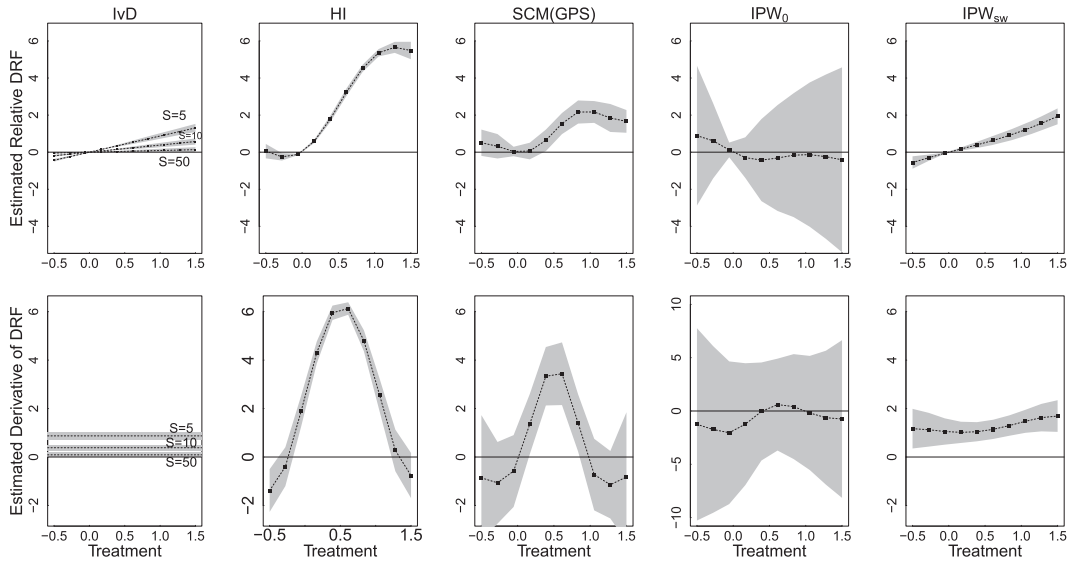


**Figure 1.** Results of Simulation Study I. The first row plots the estimated relative DRF with the horizontal solid line representing the true relative DRF. For IvD, we use $S = 5$, 10, 50 subclasses. The solid diagonal line for the method of HI is the unadjusted regression of $Y$ on $T$. The second row plots the estimated derivatives of the DRF with the solid line representing the truth. In both rows, the grey shaded areas represent 95% confidence intervals. The estimated derivative for $IPW_0$ is plotted on a different scale as its standard error is significantly larger than that of the other methods.

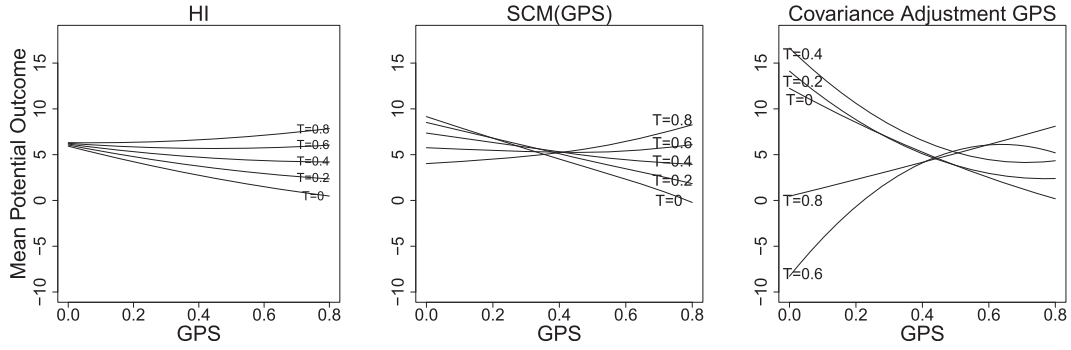


**Figure 2.** The Varying Flexibility of the Response Models. The plots show the mean potential outcome as a function of the GPS and $T$ under HI's quadratic response model (left panel), fitted SCM(GPS) (middle panel), and covariance adjustment GPS (right panel). Covariance adjustment GPS fits a quadratic regression ($Y \sim R + R^2$) in each of several subclasses based on $T$, see online Appendix 2 for details. We use 10 subclasses but only plot five. Subclassification is by far the most flexible of the three response models.

One reason that GPS-based methods can perform poorly is that their response model are based on overly strong parametric assumptions, especially equation (3). This is illustrated in Figure 2 which compares the fitted mean potential outcome as a function of the GPS and $T$ under the HI model (left panel) and under SCM(GPS) (middle panel) (online Appendix 1.1 provides additional discussion based on analytical calculations.) The fitted potential outcomes differ substantially and are considerably more constrained under the quadratic model of HI. To fit an even more flexible response model, we subclassified the data into 10 subclasses based on $T$, and fit a quadratic regression for $Y$ as a function of the GPS separately within each of the subclasses. Five out of the 10 within subclass fit are plotted in the right most plot in Figure 2. The results differ substantially from HI and reveal the considerable constraint of the quadratic response model. Subclassifying on $T$ in this way leads to a new response model and a corresponding new GPS-based estimate of the DRF; this method is discussed in online Appendix 2.

Given the relatively large sample size used in Simulation I ($n = 2000$), we replicate the study with a reduced sample size ($n = 500$). Although we observe an increase in standard errors for both the estimated relative DRF and

its derivative, the qualitative relative performance of the methods remains the same; results appear in online Appendix 1.2.

## 3.2 Simulation study II

Although IvD's response model in Simulation I is misspecified in terms of its adjustment for $X$, the method may benefit from its assumption that the DRF is linear in $T$. We address this in the second simulation that compares the performance of the methods under several alternative generative models. We also explore the frequency properties of the methods.

Suppose we have a simple random sample of 2000 observations that includes a trivariate normal covariate, $(X_1, X_2, X_3)$, with mean vector (1, 1, 4), component variances all equal to one, $\text{Corr}(X_1, X_2) = 0.3$, $\text{Corr}(X_1, X_3) = -0.4$, and $\text{Corr}(X_2, X_3) = 0.6$. Suppose further that the treatment is generated according to $T|\boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(X_1 - X_2^2 + 0.5X_3, 1)$ and the response is generated according to one of four response models

- **Gaussian Linear DRF:** $Y(t)|t, \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(X_1 + 2X_2 + t, 9)$,
- **Gaussian Quadratic DRF:** $Y(t)|t, \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}((X_1 + 2X_2 + t)^2, 9)$,
- **Lognormal DRF:** $Y(t)|t, \boldsymbol{X} \overset{\text{ind}}{\sim} \exp\left\{\frac{\mathcal{N}(X_1+2X_2+t,9)}{5}\right\}$, or
- **Gaussian Piecewise-Quadratic DRF:** $Y(t)|t, \boldsymbol{X} \overset{\text{ind}}{\sim} \begin{cases} \mathcal{N}((X_1 + 2X_2)^2, 9) & t \leq 0 \\ \mathcal{N}((X_1 + 2X_2 + t)^2, 9) & t > 0 \end{cases}$

Unlike the Gaussian response models the lognormal model exhibits significant heteroskedasticity; the central 95% of the conditional variances range from 0.3 to 17.8. On the other hand, the piecewise-quadratic model shows the greatest degree of nonlinearity.

To isolate the difference between the methods, we use the correctly specified treatment model in our analyses. (We use the R function density to estimate the marginal distribution of $T$ for use in stabilized weights.) For the fitted response model under HI and IvD's methods, we consider Gaussian regression models that are linear and quadratic in $T$. In particular for the method of IvD we fit (i) $Y \sim T$ and (ii) $Y \sim T + T^2$ within each of $S = 10$ equally sized subclasses, and for the method of HI we fit (i) $Y \sim T + R + R^2 + R \cdot T$ and (ii) $Y \sim T + T^2 + R + R^2 + R \cdot T$. With IvD, the relative DRF is computed by averaging the coefficients of the within subclass models. The response models given in Table 2 are used for SCM(GPS), $\text{IPW}_0$, and $\text{IPW}_{\text{SW}}$. For the GPS-based methods and IPW, the relative DRF is evaluated at 10 equally spaced values of $t$ between $-1.9$ to 3.4. The entire procedure was repeated using all methods on each of 1000 data sets generated with the same covariate and treatment models and with each of the three generative response models. All of the fitted response models are misspecified in their adjustment for $X$ and/or $T$, as we expect in practice. Thus, this simulation study investigates the robustness of the methods to typical misspecification of the response model.

Figures 3 and 4 report the average of the estimated relative DRFs across the simulations (dashed lines) along with their two standard deviation intervals (shaded regions). The true relative DRF functions are plotted as solid lines.

The IvD method performs reasonably well when the generative model is linear (row 1). When IvD is fit with a quadratic model (column 3), the DRF is estimated with little bias though the estimate has higher variability. IvD exhibits significant bias only when the fitted model is linear and the true DRF is non-linear (column 1, rows 2 and 3). The HI method, on the other hand, exhibits appreciable bias even when the fitted response model matches the true model in its functional dependence on $t$. Like the IvD method, the bias is most acute when the fitted model is linear but the true DRF is not. Unlike with IvD, however, the 95% frequency intervals of HI miss the true value appreciably across a wide range of treatment values.

The first three columns of Figure 4 give result for SCM(GPS), $\text{IPW}_0$, and $\text{IPW}_{\text{SW}}$ and show that $\text{IPW}_0$ and $\text{IPW}_{\text{SW}}$ improve on HI for both the quadratic and piecewise-quadtratic DRF, although the variances are larger. For the linear DRF, SCM(GPS) is comparable to HI while $\text{IPW}_0$ exhibits enormous variance, and $\text{IPW}_{\text{SW}}$ exhibits moderate variance. Except for the variance of $\text{IPW}_0$, the heteroscedasticity of the lognormal response model does not significantly effect the three methods.

As in Simulation I, we replicate Simulation II with a reduced sample size of $n = 500$. Although standard errors are larger with the smaller sample size, the relative performance of the methods is qualitatively similar; see online Appendix 1.2.

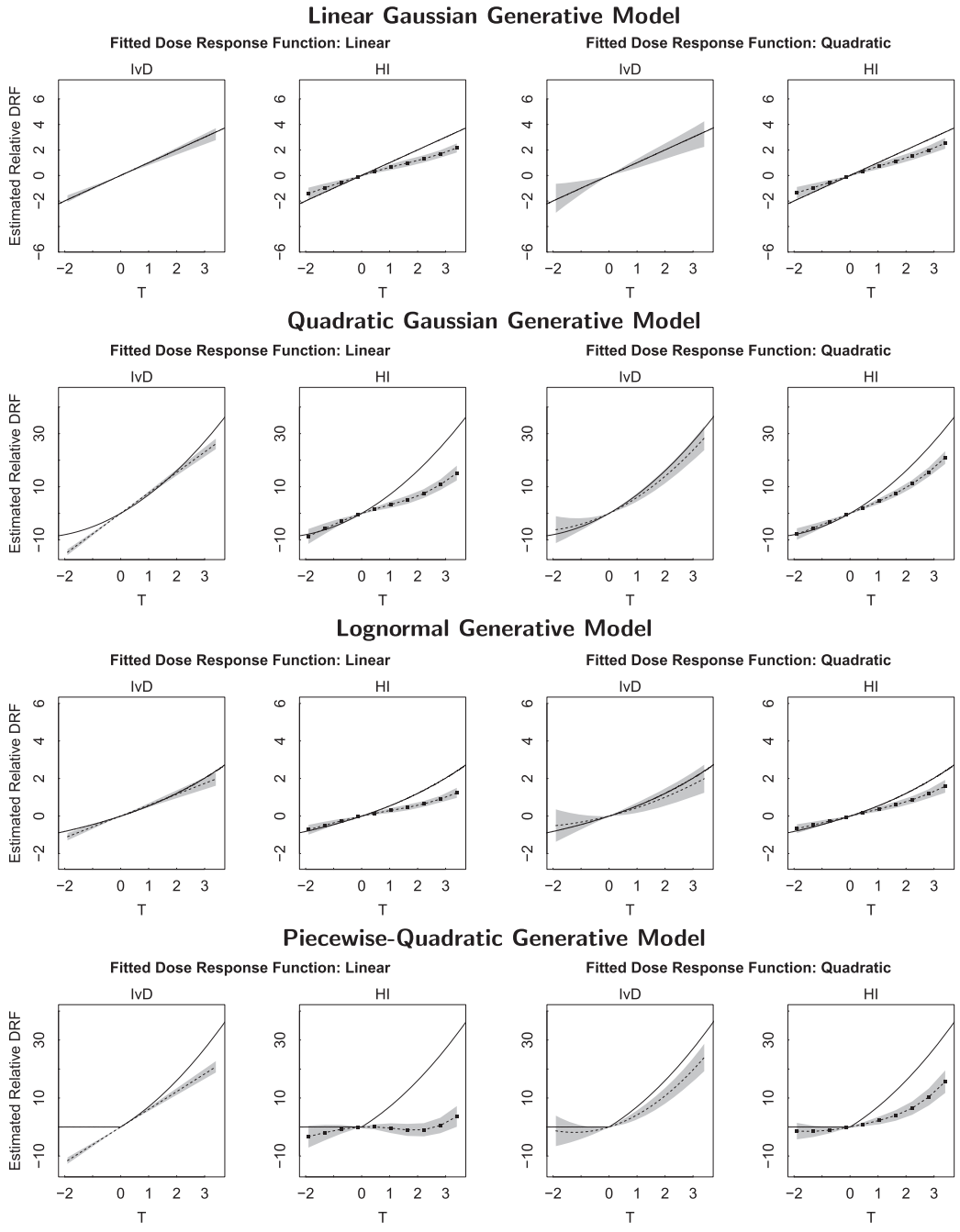


**Figure 3.** Estimated Relative DRFs in Simulation Study II Using the Methods of IvD and HI. The solid lines plot the true relative DRFs, the dashed lines plot the means of the fitted relative DRFs across 1000 simulations, and the gray shaded regions plot two standard deviation pointwise intervals across the 1000 fits. The evenly spaced grid of evaluation points used with HI is also plotted as solid circles. The method of HI shows a substantial bias with all six combinations of generative and fitted response models. The method of IvD, on the other hand, exhibits significant bias only when the fitted model is linear and the generative is not.

## 3.3 Simulation study III

This simulation study aims to investigate the effect of a heteroscedastic treatment and the robustness of the methods to misspecification of the treatment model. The simulation setup is exactly as in Simulation study II (i.e. the same sample size, covariate distribution, and replication) but we consider a heteroscedastic treatment,

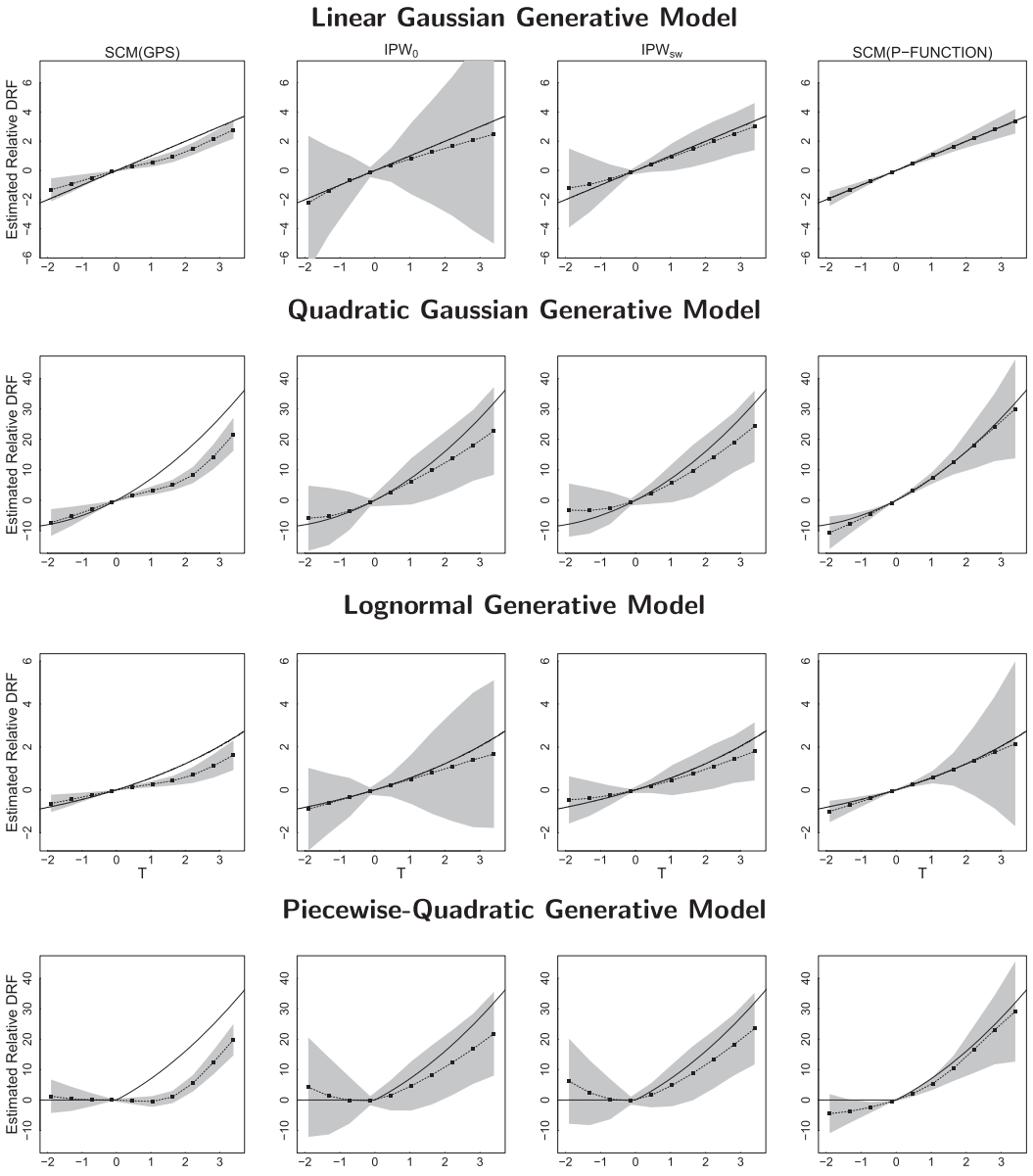


**Figure 4.** Estimated Relative DRFs in Simulation Study II for the SCM(GPS), $IPW_0$, $IPW_{SW}$, and SCM(PF) Methods. Solid lines, dashed lines, and gray regions represent the true relative DRFs the means of the 1000 fitted relative DRFs and 95% pointwise intervals. Points represent the evenly spaced grid points. The SCM(PF) method is discussed in Section 4.1.
Linear Gaussian Generative Model
Quadratic Gaussian Generative Model
Lognormal Generative Model
Piecewise-Quadratic Generative Model

$T|\boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(X_1 - X_2^2,\, 0.25X_3^2)$. The response is generated according to the Gaussian quadratic DRF of Simulation study II. This simulation is repeated using two fitted treatment models

- **Correctly Specified Treatment Model:** $T|\boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_1 + \beta_2 X_2^2,\, \sigma^2 X_3^2)$
- **Misspecified Treatment Model:** $T|\boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_1 + \beta_2 X_2^2,\, \sigma^2)$.

(We consider misspecification of both the mean and variance of the treatment model in section 5.2.) The parameters of both treatment models are fit via maximum likelihood and the marginal treatment model is fit

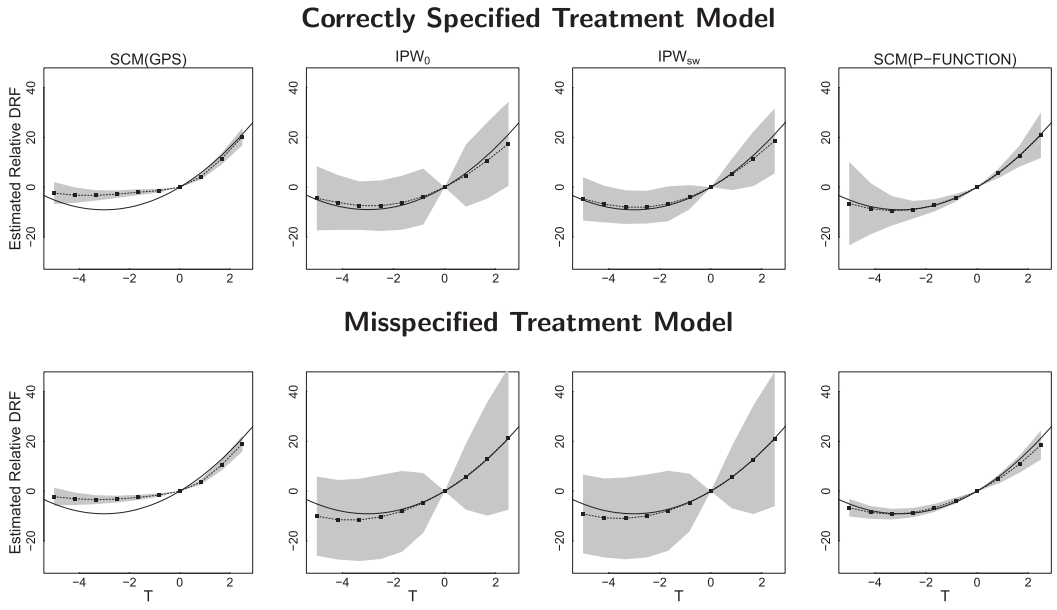


**Figure 5.** Estimated Relative DRFs under the Heteroscedastic Treatment of Simulation Study III. Solid lines, dashed lines, and gray regions represent the true relative DRFs, the means of the 1000 fitted relative DRFs and 95% pointwise intervals.
Linear Gaussian Generative Model
Quadratic Gaussian Generative Model
Lognormal Generative Model
Piecewise-Quadratic Generative Model

using the R function density. The estimated DRFs fit with SCM(GPS), $IPW_0$,[c] and $IPW_{SW}$ using the response models in Table 2 appear in the first three columns of Figure 5. The specification of the treatment model has little effect on SCM(GPS). For both of the IPW methods, the choice of treatment model effects the variance more than the bias of the fitted DRF.

Taking the results of Simulations I to III together, $IPW_{SW}$ preforms better than the other methods. The GPS methods (HI, SCM(GPS), and $IPW_0$) are explicitly designed to estimate the DRF but seem ill-suited to the task.

## 3.4 Theoretical considerations and methodological implications

Both the GPS and PF generalize RR's propensity score. In the binary case, $p(T|\boldsymbol{X})$ is uniquely determined by $e(\boldsymbol{X}) = p(T = 1|\boldsymbol{X})$. IvD focus on *uniquely* determining the full conditional distribution of $T$ given $\boldsymbol{X}$, and assume this conditional distribution is parameterized and can be uniquely represented by $\boldsymbol{\theta}$. HI, on the other hand, does not constrain the treatment assignment model in this way and instead, following the binary propensity score, evaluates $p(T|\boldsymbol{X})$ to compute propensity weights or GPS. In this way the GPS does *not* uniquely determine $p(T|\boldsymbol{X})$. There may be multiple distributions that are equal when evaluated at a particular $t$. The assumption of a uniquely parameterized propensity function constrains the choice of treatment model that can be used for a PF. In practice, however, the same treatment models are typically used by both methods.

Comparing IvD and HI, the assumption of IvD allows a stronger form of *strong ignorability of the treatment assignment given the propensity function*. In particular, Result 2 of IvD states

- **Ignorability of IvD:** $p\{Y(t)|T, e(\cdot|\boldsymbol{\theta})\} = p\{Y(t)|e(\cdot|\boldsymbol{\theta})\}$ for every $t$.

Whereas, in their Theorem 2.1, HI show

- **Ignorability of HI:** $p_T\{t|r(t,\boldsymbol{X}), Y(t)\} = p_T\{t|r(t,\boldsymbol{X})\}$ for every $t$.

In the case where $T$ is categorical, HI's ignorability implies that $\mathbf{1}\{T = t\}$ and $Y(t)$ are independent given $r(t,\boldsymbol{X})$, where the GPS is evaluated at the particular value of $t$ in the indicator function. Although achieving

conditional independence of $Y(t)$ and $T$ would require conditioning on a family of GPS, HI provide an insightful moment calculation to show how the response model described in section 2.2 can be used to compute the DRF. Nonetheless conditioning on either $R$ or $r(t, \boldsymbol{X})$, for any particular value of $t$ does not guarantee that $T$ is uncorrelated with the potential outcomes. The fact that the GPS does not constitute a single score for each individual restricts the response models that can be used. Subclassification, for example, is not feasible unless the classifying variable is low dimensional. The advantage of IvD over HI is that $Y(t)$ and $T$ are conditionally independent given the low-dimensional score, $\boldsymbol{\theta}$, enabling the use of a wide-range of response models.

## 4 Estimating the DRF Using the P-FUNCTION

In this section, we propose a new method for robust estimation of the DRF using the **PF**. Online Appendix 2 discusses another new robust GPS-based method.

### 4.1 Using the P-FUNCTION in a SCM to estimate the DRF

IvD developed the PF to estimate the average treatment effect, rather than the full DRF. Nonetheless we use the framework of IvD to compute the DRF in Simulation studies I and II (see Figures 1 and 3). The method we employ, however, is constrained by its dependence on the parametric form of the within subclass model. Practitioners would generally prefer a robust and flexible DRF, and here we propose a procedure that allows such estimation. We view this estimate as the best available for a DRF in an observational study.

We begin by writing the DRF as

$$E[Y(t)] = E[E[Y(t)|\boldsymbol{\theta}]] = E[E[Y(T)|\boldsymbol{\theta}, T = t]] \tag{10}$$

where the first equality follows from the law of iterated expectation and the second from the strong ignorability of the treatment assignment given the PF. We estimate the DRF using the right-most expression in equation (10) which we flexibly model using a SCM

$$E[Y(T)|\boldsymbol{\theta}, T = t] = f(\boldsymbol{\theta}, T) \tag{11}$$

where $f(\cdot)$ is a smooth function of $\boldsymbol{\theta}$ and $T$. In practice we replace $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ from the fitted treatment model. We approximate the outer expectation in equation (10) by averaging over the empirical distribution of $\hat{\boldsymbol{\theta}}$, to obtain an estimate of the DRF using a SCM of the PF

$$\hat{E}[Y(t)] = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}, t) \tag{12}$$

where $\hat{f}(\cdot)$ is the fitted SCM. We refer to this method of estimating the DRF as the SCM(PF) method and typically evaluate equation (12) on a grid of values of $t_1, \ldots, t_D$ evenly spaced in range of the observed treatments. Bootstrap standard errors are computed on the same grid.

Comparing equations (5) and (6) with equations (11) and (12), SCM(GPS) and SCM(PF) are algorithmically very similar. The primary difference is the choice between the PF and GPS in the response model. As we shall see, this change has a significant effect on the statistical properties of the estimates. Simply put, $\boldsymbol{\theta}$ is a much better behaved predictor variable than is $R$. When using Gaussian linear regression for the treatment model, for example, $\boldsymbol{\theta} = \boldsymbol{X}_i^{\top}\beta$, whereas $R$ is the Gaussian density evaluated at $T$. As illustrated in section 1.3, the dependence of the GPS on $t$ and the non-monotonicity of this dependence both complicate the response model and pose challenges to robust estimation.

Computing $\hat{E}[Y(t_0)]$ with equation (12) for some particular $t_0$ involves evaluating $\hat{f}(\cdot, t_0)$ at every observed value of $\hat{\theta}_i$. Invariably, the range of $\hat{\theta}$ observed among units with $T$ near $t_0$ is smaller than the total range of $\hat{\theta}$, at least for some values of $t_0$. Thus, evaluating equation (12) involves some degree of extrapolation, at least for some values of $t$. Luckily, this problem is relatively easy to diagnose with a scatter plot of the observed values of $(T_i, \hat{\theta}_i)$. The estimate in equation (12) may be biased for values of the treatment where the range of observed $\hat{\theta}_i$ is relatively small. As we illustrate in our simulation studies, however, equation (12) appears quite robust and this bias is small relative to the biases of other available methods.

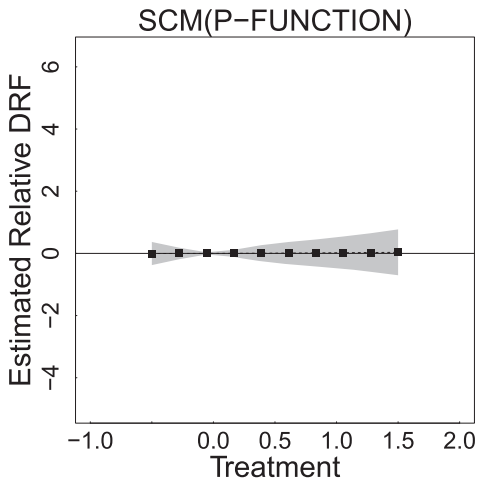


**Figure 6.** Estimated Relative DRF Using SCM(PF) in Simulation Study I. The solid (dashed) lines represent the true (fitted) relative DRF, the 95% confidence bands are plotted in grey, and the grid points are identical to those in Figure 1. The fitted relative DRF is much improved compared with those of HI, SCM(GPS), IPW0, and $IPW_{SW}$ but without the linear assumptions of IvD (see Figure 1).
Correctly Specified Treatment Model
Misspecified Treatment Model

## 4.2 Simulation studies I–III revisited

We now revisit the simulation studies from section 3, which illustrate the potentially misleading results and/or high variance of existing estimates of the DRF. Here, we compare these results with those of SCM(PF). In all cases, SCM (PF) was fitted using the same (correctly specified or misspecified) treatment assignment model and with the same equally spaced grid points. When fitting the SCM, we continue to use the penalized cubic regression spline basis for both parameters ($R$ and $T$) and a tensor product to construct a smooth fit of the continuous function $f(\theta, T)$ (see mgcv R-package documentation). Figure 6 shows the fitted (relative) DRF for SCM(PF) in Simulation I. The performance of SCM(PF) is a dramatic improvement over all other methods in Simulation I; compare Figures 1 and 6.

The rightmost column of Figure 4 presents the results of the SCM(PF) method in Simulation study II. Comparing Figures 3 and 4 again illustrates the advantages of the proposed method. The fits in Figure 3 are quite dependent on the parametric choice of the response model, whereas the non-parametric fits illustrated in Figure 4 do not require a parametric form. Among the non-parametric methods, the advantage of SCM(PF) over SCM(GPS) and $IPW_0$ is clear. It essentially eliminates bias with only a small increase in variance. Only $IPW_{SW}$ has comparable statistical properties in this simulation.

In Simulation III, the correctly specified heteroscedastic treatment model is parameterized by a bivariate function; $\boldsymbol{\theta}$ has two components which are equal to the mean and log-variance of the fitted treatment model. Both components of $\boldsymbol{\theta}$ are used as predictor variables in the SCM used to model the response. The rightmost column of Figure 5 shows that SCM(PF) performs very well in Simulation III, especially with the simpler misspecified treatment model.

Overall, SCM(PF) consistently provides better estimates than HI, SCM(GPS), and $IPW_0$, both in terms of bias and variance. $IPW_{SW}$ also preforms better than these method and is comparable to SCM(PF) in Simulation II (and Simulation IV, see online Appendix 1.3). In the bulk of our numerical studies (Simulations I and III, as well as in Simulation V in online Appendix 1.4 and the applied example in section 5), however, SCM(PF) outperforms even $IPW_{SW}$.

# 5 Example: The effect of smoking on medical expenditures

## 5.1 Background

We now illustrate the available methods by estimating the DRF of smoking on annual medical expenditures. The data we use were extracted from the 1987 National Medical Expenditure Survey (NMES) by Johnson et al.[20] Its detailed information about frequency and duration of smoking allows us to continuously distinguish among smokers and estimate the effects of smoking as a function of how much they smoke. The response variable,

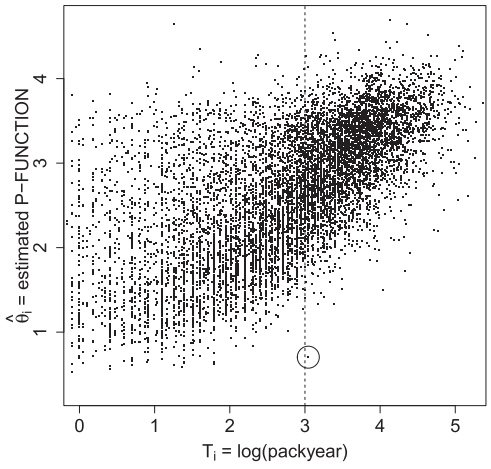


**Figure 7.** A Diagnostic for SCM(PF). Because the range of the $\hat{\theta}_i$ when $T_i > 3$ is less than the overall range of $\hat{\theta}_i$ estimating the DRF for $t > 3$ involves extrapolation under the SCM and thus possible bias. The single individual with $T_i$ slightly larger than three and $\hat{\theta}_i$ less than one is circled. Although this datapoint may mitigate bias for $t$ near three, the fitted DRF for $t > 3$ may be seriously biased.

medical costs, is verified by multiple interviews and additional data from clinicians and hospitals. IvD used the propensity function to estimate the average effect of smoking on medical expenditures. We extend their analysis and study estimation of the full DRF. Like IvD, we adjust for the following subject-level covariates: age at the times of the survey, age when the individual started smoking, gender, race (white, black, other), marriage status (married, widowed, divorced, separated, never married), educational level (college graduate, some college, high school graduate, other), census region (Northeast, Midwest, South, West), poverty status (poor, near poor, low income, middle income, high income), and seat belt usage (rarely, sometimes, always/almost always).

To measure the cumulative exposure to smoking based on the self-reported smoking frequency and duration, Johnson et al.[20] proposed the variable of `packyear`, defined as

$$\texttt{packyear} = \frac{\text{number of cigarettes per day}}{20} \times \text{number of years smoked} \tag{13}$$

We use log(`packyear`) as our treatment variable. We follow Johnson et al.[20] and IvD and discard all individuals with missing values and conduct a complete-case analysis, yielding a sample of 9073 smokers. Although in general complete-case propensity-score-based analyses produce biased causal inference unless the data are missing completely at random,[21] Johnson et al.[20] showed that accounting for the missing data using multiple imputation did not significantly affect their results.

Because the observed response variable, self-reported medical expenditure, denoted $Y$, is semicontinuous, we use the two-part model of Duan et al.[22] This involves first modeling the probability of spending some money on medical care, $\Pr(Y > 0|T, \boldsymbol{X})$, where $T = \log(\text{packyear})$, and $\boldsymbol{X}$ represents the covariates; and then modeling the conditional distribution of $Y$ given $T$ and $\boldsymbol{X}$ for those who reported positive medical expenditure. To illustrate and compare methods for computing the DRF, we concentrate on the second part of this model. Because the distribution of $Y$ is skewed, we consider the model $p(\log(Y)|Y > 0, T, \boldsymbol{X})$.

For our treatment assignment model, we use a Gaussian linear regression adjusted for all available covariates and the second order terms of two age covariates. The model was fitted using sampling weights provided with the original data. This is the same treatment assignment model used by IvD who demonstrate that it achieves adequate balance.

## 5.2 Simulation study based on the smoking data

This simulation study aims to mimic the characteristics of the actual data with the goal of comparing the statistical properties of the proposed methods in as realistic a setting as possible. In particular, we do not alter the observed covariates or treatment and use the same fitted treatment model used by IvD. Figure 7 presents a scatter plot of the observed treatment variable, $T_i = \log(\texttt{packyear})$, and the values of the PF from the fitted treatment

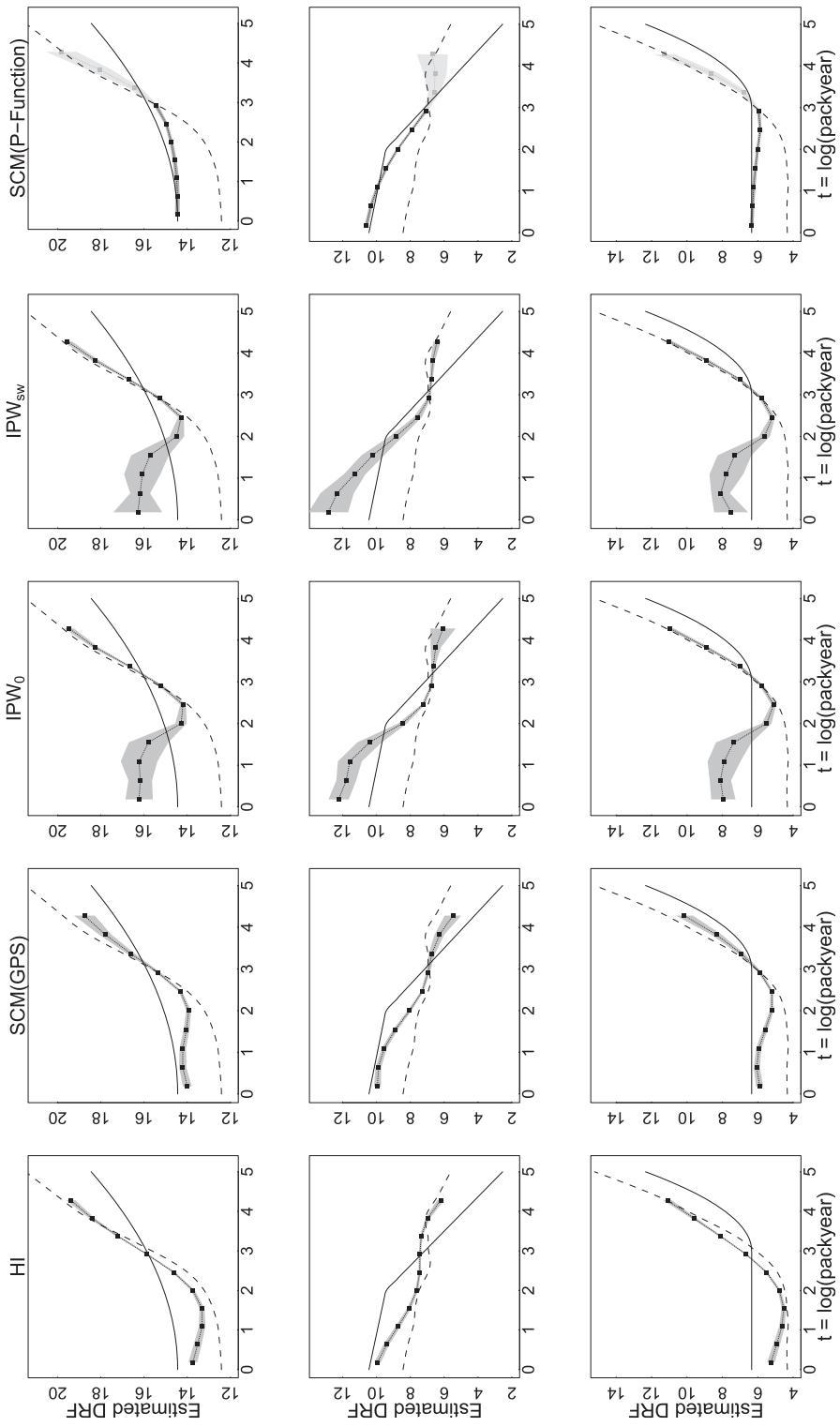


**Figure 8.** Estimated DRFs for the Simulation Based on Smoking Data. The five columns correspond to the method of HI, SCM(GPS), $IPW_0$, $IPW_{SW}$, $IPW_0$ and SCM(PF) respectively. In all plots, the dotted lines with bullets correspond to the fitted DRF. The true DRF is plotted as solid lines while dashed lines represent the fitted SCM of $\log(Y)$ on $T$, unadjusted for the covariates. The evaluation points are evenly spaced in $t$. The 95% asymptotic confidence bands plotted in grey are based on 1000 bootstrap replications. A lighter shade of grey is used in the right-most column for $t > 3$ because the estimate is less reliable in this region. The performance of the SCM(PF) clearly dominates the other methods, especially for $t < 3$.

assignment model, $\hat{\theta}_i$. As discussed in section 4.1, this plot can be used as a diagnostic for SCM(PF). Recall that this estimate requires that we fit a SCM to predict the response variable as a function of $T_i$ and $\hat{\theta}_i$. To estimate the DRF at $t$, we must evaluate the fitted SCM at $(t, \hat{\theta}_i)$ using each observed $\hat{\theta}_i$ in the data set. This involves extrapolation and thus possible bias if the range of $\theta_i$ at a particular value of $t$ is less than the overall range of $\theta_i$. Judging from Figure 7, this is a concern for $t$ greater than about three. There is a solitary individual with $T_i$ slightly above three and $\hat{\theta}_1 < 1$ that is circled in Figure 7. Even this single point can guard against significant extrapolation bias for $t$ less than three, but the concern remains for larger $t$. We emphasize that this diagnostic is performed before the response model is fit.

To explore the robustness of the methods to different DRFs, we simulate the response variable under three known DRFs and attempt to reconstruct them using HI, SCM(GPS), $\text{IPW}_0$, $\text{IPW}_{\text{SW}}$, and SCM(PF). In particular, we assume $\log(Y_i(t)) \sim \mathcal{N}(E[\log(Y_i(t))], 0.5^2)$ with $t = \log(\texttt{packyear})$ and consider three functional forms for $E[\log(Y_i(t))]$:

$$\text{Quadratic DRF}: \quad E[\log(Y_i(t))] = \frac{4}{25} \cdot t^2 + [\log(\texttt{age}_i)]^2$$

$$\text{Piecewise} - \text{Linear DRF}: \quad E[\log(Y_i(t))] = \begin{cases} -4 - 0.5 \cdot t + [\log(\texttt{age}_i)]^2, & t \leq 2 \\ -5 - 2.3 \cdot (t-2) + [\log(\texttt{age}_i)]^2, & t > 2, \end{cases}$$

$$\text{Hockey} - \text{Stick DRF}: \quad E[\log(Y_i(t))] = \begin{cases} -8.1 + [\log(\texttt{age}_i)]^2, & t \leq 3 \\ -8.1 + 1.5 \cdot (t-3)^2 + [\log(\texttt{age}_i)]^2, & t > 3, \end{cases}$$

where age is the age at the time of the survey. We include age because it is the covariate most correlated with $\log(\texttt{packyear})$ and thus most able to bias a naive analysis. Each of the response models was fitted using the sampling weights.[d]

Each of the five methods was fitted to one data set generated under each of the three DRFs. We evaluate each DRF at 10 points equally spaced between the 5% and 95% quantiles of log(packyear). The results appear in Figure 8 in which rows correspond to the three generative models and columns represent the method used to fit the DRF. In all plots, the true DRF is plotted as a solid line and a directly fitted SCM of $\log(Y)$ on $T$ as a dashed line. This SCM fit is a simple bench mark; it does not account for covariates in any way, in particular it does not adjust for any summary of the treatment assignment model. The fitted DRFs are plotted with dotted lines and bullets indicate the grid where the estimates are evaluated. The shaded regions represent 95% point-wise bootstrap confidence intervals. The diagnostic described in Figure 7 indicates possible bias in the SCM(PF) method for $t > 3$. Thus, we plot the fit in this region in light grey to emphasize its potential unreliability.

The HI fit misses the true DRF under all three generative models, even the quadratic DRF which coincides with the parametric dependence of $\log(Y)$ on $T$ under HI's response model. Instead, HI's fitted DRF tends to follow the unadjusted SCM fit of $\log(Y)$ on $T$. Although SCM(GPS) improve somewhat on HI, it still exhibits a cyclic pattern; notice its cubic-like fits in the first and third rows. Unfortunately, $\text{IPW}_0$ again exhibits instability, although in this case it takes the form of bias rather than variance. Although $\text{IPW}_{\text{SW}}$ performs significantly better than $\text{IPW}_0$ in our simulation studies, here the methods are essentially indistinguishable. Finally, SCM(PF) closely matches the true DRF under all three generative models, at least for $t < 3$. As discussed above, we suspect bias for $t > 3$ and see that the fitted DRF reverts to the unadjusted SCM in this range. The quality of the fit can be improved still further by increasing the dimension of the basis used in the SCM. We do not purse this strategy, however, for fear of over fitting. Overall, SCM(PF) appears to be the most reliable, especially considering the diagnostic that alerts us the ranges of $t$ where there is the potential for bias.

## 6 Concluding remarks

Propensity score methods have gained wide popularity among applied researchers in a number of disciplines. Although they were originally designed exclusively for binary treatment regimes, the fact that treatment variables of interest are not binary in many research settings has led to proposals for generalized propensity score methods. These methods are applicable to a variety of non-binary treatment regimes, and their applications are becoming increasingly common.

In this article, we compare two frequently used generalized propensity score methods, the GPS of Hirano and Imbens[2] and the PF of Imai and van Dyk,[3] as well as the GPS extensions of Flores et al.[8] First, we show that the suggested implementation of the HI method is sensitive to misspecification of the response model. Second, we show that while SCM(GPS) exhibits substantial improvement over HI's method, it remains biased and/or can exhibit a cyclic artifact in some situations. Third, we demonstrate that while Flores et al.'s $IPW_0$ can be highly unstable, using RHB's stabilized weights can significantly improve its performance. Finally, while IvD provides a relatively robust estimate of the average causal effect, its main limitation is its inability to estimate the DRF. We show how to obtain an estimate of the DRF based on the PF and empirically compare its performance to that of the other estimates. We also give an explanation as to why the SCM(PF) method outperforms the SCM(GPS) method. While SCM(PF) preforms well in comparison to other methods, it remains biased in realistic settings. *We emphasize that researchers should be cautious when using any method to estimate the full DRF with a continuous treatment in an observational study*.

There are several important challenges that must still be addressed. We have largely assumed that the propensity weights, GPS, and PF can be correctly estimated. This is an optimistic assumption given that modeling a multi-valued or continuous treatment in a high-dimensional covariate space is much more difficult than doing so for a binary treatment. Diagnostic tools developed for the binary treatment case are also not directly applicable to general treatment regimes. Even more challenging is diagnosing misspecification in the response model. As we have illustrated, this can lead to significant bias in the estimated DRF. Our proposals rely on implementing more flexible response models in more natural spaces, but principled diagnostics for the response model remain elusive. Diagnosing and correcting for imbalance in either the PF or the GPS is another difficulty. Since the subpopulation that has propensity for treatment varies with the dose, the estimated dose–response function is in effect the treatment effect on a varying subpopulation. Recently, researchers have started to develop new methods for estimating the propensity weights, GPS, and PF in the presence of possible misspecification of the treatment assignment model.[6,7] Future research should also address the estimation of the DRF in the presence of possible misspecification of the response model, as well as diagnostics tools.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Kosuke Imai https://orcid.org/0000-0002-2748-1022

## Notes

a. There are different nonparametric methods based on this model. For example, FFGN propose a nonparametric kernel estimator with a polynomial regression of order 1,[19] while Bia et al.[18] propose using radial basis functions to avoid the need for tensor product splines (they also propose a penalized spline method with tensor product).[23] We use the SCM to facilitate comparisons of the methods. As with equation (2), we use the mgcv package with penalized cubic regression splines as the basis functions with dimension equal to five for both $T$ and $\hat{R}$ along with a tensor product. We leave the comparison of these different nonparametric methods to future research.

b. It would also complicate estimation of the DRF. Because the treatment assignment mechanism is strongly ignorable given the propensity function (IvD), we aim to adjust for the propensity function in a robust manner in the response model. Thus, adjusting for $\hat{\theta}$ within the subclasses poses no conceptional problem. In practice, however, $\hat{\theta}$ tends to be fairly constant within subclasses and its coefficient tends to be correlated with the intercept. A solution is to recenter $\hat{\theta}$ within each subclass. Because, we propose a more robust strategy for estimating the DRF using the PF in section 4.1, however, we do not purse such adjustment strategies here.

c. Using the bandwidth of Fan and Gijbels[19] as suggested by FFGN leads to numerical instability in 708 of the 1000 datasets fit with $IPW_0$ under the correctly specified treatment model. The bandwidth can be tuned to improve stability; results in Figure 5 are based on a single bandwidth that gave stable results in 793 of the 1000 datasets. Although this procedure would be difficult to implement with a single dataset, it gives the best possible representation to $IPW_0$. The standard bandwidth of Fan and Gijbels[19] was used with the misspecified treatment model.

d. When using $IPW_0$ or $IPW_{SW,}$ we construct new weights by multipling the weights required by IPW and the sampling weights. We also take the sampling weights into account when estimating the marginal distribution of the treatment with $IPW_{SW}$ (using the `density` function in R). Ignoring the sampling weights leads to similar results.

## References


1. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
2. Hirano K and Imbens GW. The propensity score with continuous treatments. In: *Applied Bayesian modeling and causal inference from incomplete-data perspectives: an essential journey with Donald Rubin's statistical family*. New York: Wiley, 2004, pp.73–84.
3. Imai K and van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *J Am Stat Assoc* 2004; **99**: 854–866.
4. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**: 706–710.
5. Robins JM, Hernán MA and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–560.
6. Fong C, Hazlett C and Imai K. Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements. *Ann Appl Stat* 2018; **12**: 156–177.
7. Zhu Y, Coffman D and Ghosh D. A boosting algorithm for estimating generalized propensity scores with continuous treatments. *J Causal Inference* 2015; **3**: 25–40.
8. Flores CA, Flores-Lagunes A, Gonzalez A, et al. Estimating the effects of length of exposure to instruction in a training program: the case of job corps. *Rev Economics Stat* 2012; **94**: 153–171.
9. Joffe MM and Rosenbaum PR. Propensity scores. *Am J Epidemiol* 1999; **150**: 327–333.
10. Lu B, Zanutto E, Hornik R and Rosenbaum PR.) Matching with doses in an observational study of a media campaign against drug abuse. *J Am Stat Assoc* 2001; **96**: 1245–1253.
11. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc* 1987; **82**: 387–394.
12. Hernán MA, Brumback B and Robins JM. Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Stat Med* 2002; **21**: 1689–1709.
13. Hogan JW and Lancaster T. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Stat Meth Med Res* 2004; **13**: 17–48.
14. Moodie EE and Stephens DA. Estimation of dose-response functions for longitudinal data using the generalised propensity score. *Stat Meth Med Res* 2012; **21**: 149–166.
15. Ertefaie A and Stephens DA. Comparing approaches to causal inference for longitudinal data: Inverse probability weighting versus propensity scores. *Int J Biostat* 2010; **6**: 1–24.
16. Rubin DB. Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Stat Sci* 1990; 5: 472–480.
17. Imai K, King G and Stuart EA. Misunderstandings among experimentalists and observationalists about causal inference. *J Royal Stat Soc, Ser A (Statistics in Society)* 2008; **171**: 481–502.
18. Bia M, Flores AC and Mattei A. Nonparametric estimators of dose-response functions. *CEPS/INSTEAD* 2011. Working Paper No. 2011-40.
19. Fan J and Gijbels I. *Local polynomial modelling and its applications*. London: Chapman and Hall, 1996.
20. Johnson E, Dominici F, Griswold M, et al. Disease cases and their medical costs attributable to smoking: an analysis of the National Medical Expenditure Survey. *J Econometrics* 2003; **112**: 135–151.
21. D'Agostino RB Jr and Rubin DB. Estimating and using propensity scores with partially missing data. *J Am Stat Assoc* 2000; **95**: 749–759.
22. Duan N, Manning WGJ, Morris CN, et al. A comparison of alternative models for the demand for medical care. *J Business Economic Stat* 1983; **1**: 115–126.
23. Bia M, Flores C, Flores-Lagunes A, et al. A Stata package for the application of semiparametric estimators of dose-response functions. *Stata J* 2014; **3**: 580–604.