Featured Article

Statistics in Medicine

Received 26 October 2015, Accepted 31 March 2016 Published online 3 May 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6973

A flexible, interpretable framework for assessing sensitivity to unmeasured confounding

Vincent Dorie,[a] Masataka Harada,[b] Nicole Bohme Carnegie[c] and Jennifer Hill[a*†]

When estimating causal effects, unmeasured confounding and model misspecification are both potential sources of bias. We propose a method to simultaneously address both issues in the form of a semi-parametric sensitivity analysis. In particular, our approach incorporates Bayesian Additive Regression Trees into a two-parameter sensitivity analysis strategy that assesses sensitivity of posterior distributions of treatment effects to choices of sensitivity parameters. This results in an easily interpretable framework for testing for the impact of an unmeasured confounder that also limits the number of modeling assumptions. We evaluate our approach in a large-scale simulation setting and with high blood pressure data taken from the Third National Health and Nutrition Examination Survey. The model is implemented as open-source software, integrated into the treatSens package for the R statistical programming language. © 2016 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: Bayesian modeling; causal inference; nonparametric regression; sensitivity analysis; unmeasured confounding

1. Introduction

Causal inference in the absence of a randomized experiment or strong quasi-experimental design requires appropriately conditioning on all pre-treatment variables that predict both treatment and outcome, also known as confounding covariates. This requirement, formalized as the ignorability assumption in the statistics literature, is often not satisfied, which leaves inference vulnerable to bias. Researchers interested in causal questions that cannot be addressed with randomized experiments are thus left in the unenviable position of either avoiding causal language or arguing for the satisfaction of a strong and untestable assumption. The sensitivity of a study to this ignorability assumption can be analyzed by positing the existence of an unmeasured confounder and specifying its form in the inferential model. If the treatment effect estimate under the augmented model differs substantially from the original under plausible levels of confounding, then the study can be deemed sensitive to violations of ignorability.

In addition to structural assumptions required for the causal estimand to be identifiable (i.e., ignorability), bias can also be introduced when making assumptions about the form of the causal pathway. In practice, these assumptions often take the shape of parametric models, in which the exact relationships among response, treatment, and covariates are made explicit. If this parametric form is incorrect, model misspecification biases may be introduced. However, these assumptions and attendant biases can be mitigated by employing nonparametric models. Flexible nonparametric methods can represent functional forms of arbitrary complexity so that, provided that ignorability holds, the true relationship of the

[a]*Humanities & the Social Sciences, New York University, New York, NY, U.S.A.*
[b]*Economics, Fukuoka University, Fukuoka, Japan*
[c]*Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee,WI, U.S.A.*
[*]*Correspondence to: Jennifer Hill, 246 Greene Street, Room 804 New York, NY 10003, U.S.A.*
[†]*E-mail: jennifer.hill@nyu.edu*

response variable to the treatment variable and covariates can be recovered. Causal inference can be conducted by using a fitted nonparametric model to make predictions for the counterfactuals, which are in turn used to compute estimates of treatment effects.

In this paper, we present a simulation-based approach to test the sensitivity of a study to unmeasured confounding that utilizes nonparametric methods for modeling the response variable. We explicitly include the unmeasured confounder in both the response and treatment models as an additive term with coefficients that serve as sensitivity parameters. Because the unmeasured confounders behave as latent variables, completing the model with weakly informative priors allows us to draw samples from the posterior distribution of the treatment effect using Markov chain Monte Carlo. For any treatment effect estimate, the values of the sensitivity parameters are graphically compared with the marginal effects[‡] of observed covariates, so that the researcher can have some benchmark for deciding problematic levels of confounding are plausible. The nonparametric method we use to model the response surface is called Bayesian Additive Regression Trees (BART), which has been shown to perform well in a wide variety of settings without requiring the adjustment of tuning parameters.

2. Background

There has long been dissatisfaction with relying on observational studies to answer causal questions. Many approaches have been proposed to try to reduce the dependence on standard ignorability assumptions in non-experimental work by focusing on quasi-experimental designs and natural experiments [1,2]. However, it is not always possible to find data that meet these criteria and can also address the research question of primary interest. Moreover, these approaches come with their own sets of assumptions, which are not always more plausible than the standard ignorability assumption in a typical observational study. For instance, traditional formulations of the instrumental variables approach require satisfying the following assumptions: (i) ignorability of the instrument, (ii) the exclusion restriction (colloquially, the instrument can only affect the outcome through its effect on the treatment), and (iii) the monotonicity assumption. It is relatively rare to find compelling examples of such instruments in practice, and when the instrument is weak (in the sense that there is a low percentage of observational units whose behavior is influenced by the instrument), violations of these assumptions can lead to extreme bias [3]. Finally, quasi-experimental and natural experiment approaches often yield inferences about only a small subset of the population of interest, which can be unsatisfying [4].

Another way to address concern regarding violations of the ignorability assumption is to directly assess the sensitivity of a given study to violations of the ignorability assumption. Many strategies have been proposed that explore the impact on causal estimates of the inclusion of an unmeasured confounder that, along with the observed covariates, would serve to satisfy the ignorability assumption [e.g., [5–10]]; this is often referred to as *sensitivity analysis* (SA).

While unmeasured confounding represents one source of potential bias, an over-reliance on parametric assumptions can introduce another. In practice, it can be difficult to diagnose and fix deviations from linearity and additivity required by the most common parametric models in high-dimensional space (that is, when there are many covariates). Furthermore, the iterative process of model diagnosis and model tweaking (where at each stage the researcher can see the new treatment effect estimate) can inadvertently lead to a tendency to fit models that yield treatment effects that conform to a priori beliefs about the sign and magnitude of these effects.

Figure 1 illustrates the importance of including nonlinear effects and interactions in the context of our motivating example, an investigation of the effect of medication on blood pressure. Using data from the Third National Health and Nutrition Examination Survey [NHANES III, [11]], this figure displays a locally weighted scatterplot smoothing (LOESS) fit to the expected value of blood pressure conditional on age, separately for males and females. The shaded bands depict pointwise confidence intervals for these expectations. These plots reveal a strongly nonlinear relationship between blood pressure and age, and moreover, these relationships differ between the sexes. Finding and appropriately modeling all such nonlinearities and interactions can be challenging when many covariates are required to satisfy ignorability.

[‡]*By 'marginal effect', we do not mean to imply a casual relationship - instead only that of the expected difference in response between two individuals whose covariates differ only in a single predictor, one of which is one half of a standard deviation above its mean while the other is one half of a standard deviation below.*

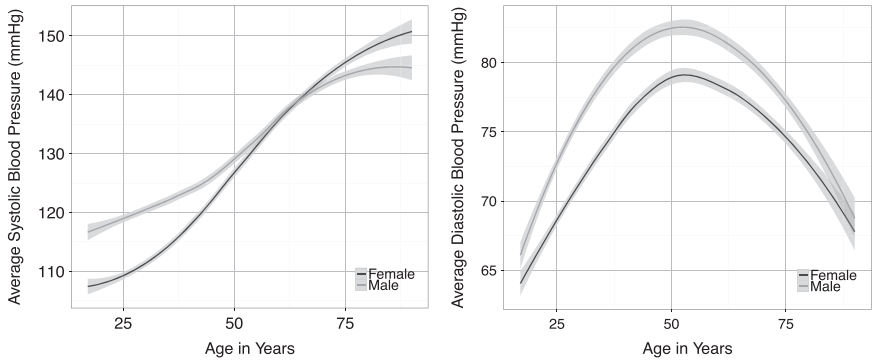

Figure 1. LOESS curves of average systolic (left) and dystolic (right) blood pressure for all individuals in the Third National Health and Nutrition Examination Survey dataset, plotted against age and separated by sex. The shading shows pointwise 95% confidence intervals for the mean.

2.1. Existing approaches to sensitivity analysis

Modern approaches to SA can be categorized according to the number and type of their sensitivity parameters. Sensitivity parameters are the values that control how the unmeasured confounder enters into the model and must be interpretable by the researcher to be useful. An important example is Rosenbaum's Γ, which bounds the odds that one member of a matched pair receives the treatment relative to the other [12]. Working with a single parameter, so-called primal methods, specifies the relationship between the unobserved confounder and the treatment assignment mechanism but assume that the confounder and response are essentially collinear [e.g., [13]]. Conversely, 'dual' methods assume the inverse set of relationships [e.g., [9]]. As we are primarily motivated by relaxing assumptions, we specify both relationships and consequently define two sensitivity parameters. Methods of this type are sometimes classified as 'simultaneous' [e.g., [14, 15]]; however, we will use the term 'two parameter'.

Many one-parameter (primal and dual) SA approaches have the advantage of being nonparametric or semi-parametric. Of these methods, the majority rely on randomization tests, such as McNemar's test for binary treatment and response [13] or the Wilcoxon signed-rank test for a continuous response [13, 16]. An overview of such approaches to SA can be found in Chapter 4 of [10]. Unfortunately, these SA procedures have been shown to be sensitive to the choice of test statistic [17]. In addition, many of these methods also require matched samples, a complicating factor that we will discuss later. Finally, one-parameter SA approaches have the disadvantages of using sensitivity parameters that are not always easily interpretable and of reliance on overly conservative assumptions, such as the assumption in primal methods that the unobserved confounder is nearly perfectly correlated with the outcome variable.

Two-parameter SA approaches, on the other hand, tend to have more interpretable parameters (expressed as partial correlations or regression coefficients), do not require matching, and do not require assumptions about strong/perfect correlations between the unobserved confounders and either the treatment or response. The trade-off, however, is that they tend to rely more strongly on parametric assumptions.

For example, in an approach proposed by [14], the response surface (expected value of the response as a function of both the confounding covariates and treatment variable) and treatment assignment mechanism (expected value of the treatment assignment as a function of the confounding covariates) are modeled using linear regression and logit models, respectively. An unmeasured confounder is assumed to exist and is added to each model, parameterized by partial correlations. The model is then fit using marginal maximum likelihood. Reference [18] relies on a similar model but uses a computationally intensive simulation-based approach to explore the treatment effect estimates that manifest across a range of sensitivity parameters. Reference [19] also uses a simulation approach but reparameterizes Imbens's model so that the sensitivity parameters can be expressed as regression coefficients, with the aim of building an easily interpretable framework that could be adapted to more complicated models. Moreover, the authors extend the framework to accommodate estimation of a wider range of estimands, such as the effect of the treatment on the treated and the effect of the treatment on the controls.

We know of two other semiparametric two-parameter sensitivity nalyses, Rosenbaum and Silber [20] and Ichino *et al.* [21]. The first - an extension of Rosenbaum's Γ - is limited by the fact that it discards information about the magnitude of the treatment effect by dichotomizing the difference in outcomes

between groups as simply positive or negative. The second uses propensity score matching techniques, which can be problematic for reasons which we discuss below.

2.2. Approaches to nonparametric or semi-parametric causal inference

Many options exist to reduce the reliance of causal inferences on parametric assumptions. For instance, conditional on the ignorability assumption, a popular conceptual approach is to find appropriate comparisons between observations through matching or weighting [[22, 23], respectively]. However, these approaches require their own restrictions, which may lead to further biases if unmet. For example, matching depends on assumptions about when sufficient balance and overlap exist, and it has been shown that the same balance definition can sometimes lead to a wide variety of potential treatment effect estimates depending on the matching procedure used [24]. On the other hand, weighting using propensity scores either relies heavily on the estimate of the propensity score, requiring its own modeling assumptions and introducing its own biases [25] or, like matching, relies instead on balance metrics to ascertain that an appropriate pseudo-comparison group has been created. The motivation behind these methods is that if well-balanced comparison group can be created, then inference should be fairly robust to misspecification of the model used to estimate treatment effects (the response surface).

As an alternative to this line of reasoning, nonparametric regression methods attempt to model the totality of the response surface, from which counterfactuals can be imputed and causal estimates calculated directly. The theory for this dates (at least) to [26], but the introduction of flexible methods to handle arbitrary complexity – heavily influenced by machine learning – is a relatively new development. These approaches have been shown to be preferable to many popular matching and weighting approaches in several scenarios [27–29].

A related concern with parametric assumptions is the growing awareness that not only is functional form important, but also the structural relationships between covariates can have an impact. For example, there is an ongoing debate about the dangers of including instrumental variables in response models, which may serve to amplify the bias because of an unobserved confounder [30, 31]. In large observational studies, the correct role for any particular variable is not always apparent. With this in mind, some researchers are opting to keep parametric models but be more sophisticated in their application by analyzing covariates before incorporating them. For example, [32] shows how the inclusion of different kinds of covariates in a propensity score model subsequently influences the treatment effect estimate by varying their relationships to the treatment and response variables. Continuing in this vein, [33] develop a complex algorithm to determine which predictors to use in high-dimensional propensity score model, which is then fit using ordinary logistic regression. On the topic of variable selection, there is conflicting advice with some arguing that the strongest estimate of the propensity score uses all available covariates [30], while others point out that aggressively balancing on observed covariates may produce an imbalance on any that are unobserved [34]. Although nonparametric methods may help address these concerns, variable selection techniques for causal inference are beyond the scope of this paper.

2.3. Paper overview

The goal in this paper is to develop a sensitivity analysis framework that is both easily interpretable but also widely applicable. These two, often competing, considerations have driven our choice of methodology. Thus, we focus on a two-parameter sensitivity analysis approach similar to [19] and use sensitivity parameters that take the form of regression coefficients (which the majority of researchers are comfortable interpreting). We embed this model within a Bayesian framework, which is fit via a Markov chain Monte Carlo sampler. This allows us to replace model components with ones that are applicable to different kinds of data or are more broad in scope in a way that may not always be tractable using marginal maximum likelihood. In particular for this paper, it allows us to swap in a Bayesian nonparametric algorithm to flexibly fit the response surface portion of the model. Taken together with a parametric assignment mechanism model, this produces a semi-parametric sensitivity analysis. An open-source implementation of the software has been added to the publicly available treatSens package [35] for the R statistical programming language [36] and is available on the Comprehensive R Archive Network. We will refer to the original algorithm henceforth as linear treatSens to distinguish it from the semi-parametric treatSens developed in this paper.

This paper proceeds by first reviewing causal inference notation, defining relevant estimands and discussing the requisite structural and parametric assumptions. We then describe the two-parameter sensitivity analysis framework defined in [19]. Subsequently, we extend that approach to allow for nonparametric specifications of the response surface and propose a specific estimation strategy involving the

BART [37] algorithm. To assess the performance of our approach, we conduct a large-scale simulation study. Finally, we illustrate its use in an applied example that estimates the effect of anti-hypertensive drugs on blood pressure using data from NHANES III.

3. Causal inference notation and assumptions

We follow convention in the statistics literature [38] by defining an individual-level causal effect of a binary treatment, Z, as a comparison across potential outcomes such as $Y_i(1) - Y_i(0)$, where $Y_i(0)$ is the outcome that would manifest for person i if $Z_i = 0$ and $Y_i(1)$ is the outcome that would manifest for person i if $Z_i = 1$.

In this notation, treatment effects are simply averages of these individual-level causal effects across subpopulations of interest. For example, the average treatment effect (ATE) is the expected value $\mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$, or the difference between the average response when everyone is treated and the average response when no one is treated (the subscript 'i' has been dropped for simplicity). Two conditional average treatment effects that are often of interest are the average treatment effect on the treated (ATT) and the average treatment effect on the controls (ATC), given by $\mathrm{E}[Y(1) - Y(0) \mid Z = 1]$ and $\mathrm{E}[Y(1) - Y(0) \mid Z = 0]$, respectively. Note that for the ATT (ATC), the difference in potential outcomes is averaged *only over those units observed to be in the treatment (control) group*.

Because we cannot observe $Y(1)$ for observations assigned to control and we cannot observe $Y(0)$ for observations assigned to treatment, these treatment effects are not identified without further assumptions. The most common assumption invoked is the so-called ignorability assumption [38], also known in various disciplines as 'selection on observables,' 'all confounders measured,' 'exchangeability,' the 'conditional independence assumption,' and 'no hidden bias' [[10, 39–41]]. A special case of the ignorability assumption occurs in a completely randomized experiment in which $Y(0), Y(1) \perp Z$. One implication is that $\mathrm{E}[Y(a) \mid Z = a] = \mathrm{E}[Y(a)]$, allowing identification of the previous estimands solely from observed outcomes.

In the absence of a randomized experiment, identification can be achieved by appropriately conditioning on the vector of confounding covariates, X, that satisfies the more general form of the ignorability assumption, $Y(0), Y(1) \perp Z \mid X$. This assumption allows us to identify average treatment effects such as the ones described earlier because, while $\mathrm{E}[Y(a) \mid Z = a] \neq \mathrm{E}[Y(a)]$, $\mathrm{E}[Y(a) \mid Z = a, X] = \mathrm{E}[Y(a) \mid X]$.

In this situation, the ATE is found by averaging the conditional expectation $\mathrm{E}[Y(1) - Y(0) \mid X] = \mathrm{E}[Y(1) \mid Z = 1, X] - \mathrm{E}[Y(0) \mid Z = 0, X]$ over the distribution of X. To obtain the ATT (or ATC), this averaging is performed over the distribution of X for the treatment (or control) group. Much of the focus of the causal inference literature in the past few decades has been on appropriate ways to estimate these conditional expectations without making strong parametric assumptions, as discussed in more detail in Section 2. This paper is similarly motivated.

4. Sensitivity analysis frameworks and assumptions

To test the sensitivity of a result to a potential unmeasured confounder, it is standard to hypothesize that such a confounder exists and determine the level of confounding required to drive the naïve treatment effect (the treatment effect estimated in the absence of this confounder) to zero or nonsignificance. In a classic early example [5], the authors quantify how implausibly strong the level of confounding created by a latent genetic factor would need to be to fully explain the association between smoking and lung cancer.

4.1. Standard formulation of sensitivity analysis

Formally, our SA proceeds by supposing that ignorability is satisfied with the addition of a confounder, U. That is, we assume that $Y(0), Y(1) \perp Z \mid X, U$. The complication is that we do not, of course, observe U and it could take any of an infinite variety of forms. However, if we specify a joint model for our observed data and U, then we can calculate how conditioning on U would change the estimated treatment effect. By comparing various manifestations of U with the observed covariates in our dataset, we can also evaluate the plausibility that such an omitted confounder exists.

We use the parametric two-sensitivity-parameter model presented in [19] as a foundation. Specifically, the original model for binary treatment variables underlying linear treatSens is as follows:

$$Y \mid X, U, Z \sim \mathrm{N}\left(X\beta^{y} + \zeta^{y}U + \tau Z, \sigma_{\mathrm{y}}^{2}\right), \tag{1}$$

$$\begin{aligned} Z \mid X, U &\sim \text{Bernoulli}\left(\Phi(X\beta^{z} + \zeta^{z}U)\right), \\ U &\sim \text{Bernoulli}(\pi^{u}), \end{aligned} \tag{2}$$

where Φ denotes the standard normal cumulative distribution function or probit link. The unmeasured confounder U is assumed to be independent of the measured confounders X. This can be conceptualized by considering U to represent the portion of the unobserved covariate not explained by observed covariates. Conveniently, the sensitivity parameters, ζ^{y} and ζ^{z}, are easily interpretable as regression coefficients from a linear regression and probit regression, respectively.

The unmeasured confounder U is assumed to be independent of the measured confounders X. This can be justified by considering U to represent the portion of the unobserved covariate not explained by observed covariates. The fact that U is specified as binary represents a limitation of the current model although one could conceptualize a latent continuous unobserved confounder with a pertinent cutoff that would map to this binary variable. Specifying U as continuous substantially increases the mathematical complexity of our fitting algorithm, however, and thus will be reserved as a topic for future work.

The algorithm proceeds by determining ranges of sensitivity parameters, ζ^{y} and ζ^{z}, that inform the treatment effect estimate. This is carried out by following the line $\zeta^{y} = \zeta^{z}$ through the point of no-significance and out until the treatment effect is approximately 0. A box that encapsulates this point will contain most, if not all, of the pairs of sensitivity parameter values where the substantive nature of the estimate would be changed. The algorithm then divides these ranges into a grid and, for each unique parameter combination, simulates values of U from its conditional distribution given the data. The estimate of the treatment effect conditional on that manifestation of U is then computed. The results can then be displayed using a contour plot to reveal the combinations of sensitivity parameters that yield various treatment effect estimates. The parameter values can further be compared with the magnitude of associations of observed confounders in the model (with non-dichotomous confounders standardized to have mean 0 and variance 1), as all terms are regression coefficients. This approach provides the foundation for an easily interpretable, two-parameter SA.

5. Extensions to the original model

We extend the formulation in [19] in two ways. First, we allow for a nonparametric fit of the response surface. Second, we create a fully Bayesian version of the model. These extensions will be motivated and discussed in more detail in this section.

5.1. Extension 1: Nonparametric Fit

Unbiased estimation of causal estimands requires not only that we have observed all the relevant confounding covariates (the ignorability assumption) but also further that we can accurately recover the relevant conditional expectations. However, the strict linearity and additivity assumptions implicit in the original formulation of the model are not always believable. Fortunately, parametric assumptions of this sort are easier to relax than structural assumptions such as ignorability.

We capitalize on recent approaches to causal inference that directly and flexibly fit the response surface and modify Equation (1) such that

$$Y \mid X, U, Z \sim \mathrm{N}\left(f(X, Z) + \zeta^{y}U, \sigma_{\mathrm{y}}^{2}\right),$$

where $f(\cdot)$ is allowed to be an arbitrary function. Although U enters the equation linearly and additively, the linear restriction is unimportant because U has been specified as a binary variable. In theory, U could enter non-additively; however, short of a very precise model for this non-additivity, doing so would require additional sensitivity parameters, which would in turn complicate interpretability.

We propose to fit this part of the model using an algorithm called BART [37] that has been demonstrated to perform well in causal inference settings [24, 27, 28, 42]. The exact method by which f is estimated is of less importance than that it can flexibly capture dependencies among covariates and between the covariates and treatment variable. While nonparametric methods have been widely applied to difficult regression problems – including such techniques as as generalized additive models [43], Gaussian processes [44], or kernel regression techniques [45] – complications arise in adapting these methods to causal

inference through the introduction of the treatment variable. For instance, in generalized additive models and Gaussian processes, one must explicitly define which terms interact with Z by choosing the additive terms or covariance function, respectively. In light of these difficulties, few nonparametric response surface models have been proposed for causal inference problems; one exception is [29]. While these issues can be addressed by direct involvement of the analyst, we prefer to utilize methods, which require the minimum of expert intervention and have a proven track record.

Not only does BART perform well at its default settings for a wide variety of causal inference problems, but it also scales well without requiring approximation techniques, has a public software implementation [46,47], and is a proper Bayesian model that can be embedded in our framework through the introduction of a posterior sampler. We discuss the specifics of the new joint BART and SA algorithm in the following section. The resulting semi-parametric treatSens algorithm is publicly available in the treatSens [35] package for the R statistical programming language [36].

5.2. *Extension 2: Fully Bayesian model*

To fully account for our uncertainty about our parameters, we create a Bayesian version of the model. Specifically, we replace the response model of linear treatSens, that is, Equation (1), with

$$Y \mid U, \mu_{xz}, \sigma^2_y \sim \mathrm{N}\left(\mu_{xz} + \zeta^y U, \sigma^2_y\right),$$
$$\mu_{xz}, \sigma^2_y \sim \mathrm{BART}(X, Z).$$

Here, BART(X, Z) signifies that Metropolis jumps for these parameters are handled externally by BART and μ_{xz} is the prediction at point (X, Z). A brief overview can be found in the next section; for full details, see [46]. Further, we impose a prior on the coefficients in the model for Z, β^z in Equation (2); we provide options for either a flat, normal, or Student-t distribution. This yields a fully Bayesian formulation, which we fit by writing a posterior sampler. To provide a point of comparison with previous methods, we also implement a variant that uses maximization for the parameters in the assignment mechanism. This algorithm falls in the class of stochastic expectation–maximization (S-EM) procedures [48]. In this case, we omit the prior for β^z.

5.3. *Bayesian Additive Regression Tree model*

Bayesian Additive Regression Trees is a sum-of-trees model that adds together the predictions of a number of regression trees suitably regularized by prior distributions. Regression trees constructed by BART partition the space of covariates by using sequential binary decision rules, each one of which splits using a single covariate. For example, when predicting blood pressure, the root of the tree might divide the observations by sex. The male subjects might be further separated into those greater than and less than or equal to 45 years old, while the female subjects sorted similarly using a different variable. The possible splits are derived from observed data according to a pre-specified rule, such as at percentiles or percentages of the distance between the smallest and largest value of a covariate.

The leaf nodes at the end of the tree contain distinct subsets of the observations defined by the covariates, and the fit for that leaf is an average of the outcomes for that leaf shrunk according to a prior distribution so as to avoid overfitting. BART is 'additive' as the predictions from many small trees ('weak learners') are summed together.

The likelihood specified by BART uses the predictions from these trees as a mean function. Observations are assumed to be independently and normally distributed about this mean and share a common variance. The model is made Bayesian by the addition of priors over the model components, namely, the space of trees, the variance term, and the mean parameters that are used in every leaf node.

We can write down a BART model succinctly as follows. Let $\mathcal{T}$ be the set of all non-empty binary decision trees that partition the values of $X_1, \ldots, X_n$ according to the method described earlier. For each of $T_1, \ldots, T_K \in \mathcal{T}$ trees, let A_{jk} be the $1, \ldots, J_k$ sets of the partition corresponding to the leaf nodes of tree k. The leaf-node parameters of tree T_k are collected in the set M_k, whose members we denote μ_{jk} and are indexed similarly. Then

$$Y_i \mid T, M, \sigma^2, X_i \overset{\text{ind}}{\sim} \mathrm{N}\left(\sum_{k=1}^{K}\sum_{j=1}^{J_k} \mu_{jk}\mathrm{I}_{\{X_i \in A_{jk}\}}, \sigma^2\right), \quad \text{for } i = 1, \ldots, n.$$

This is completed by priors $p(T)$, $p(\mu)$, and $p(\sigma)$. I_A is the indicator function of the set A.

5.4. *Choice of model for assignment mechanism*

While BART has superior properties for fitting continuous responses, its performance for binary response data can strongly depend on the choice of hyperparameters. In short, the amount of prior-assumed variability in the underlying and unconstrained function can cause the algorithm to overfit when the covariates are weakly correlated with the reponse and underfit in the converse case. A more detailed discussion is available in the Web-based Supporting Materials at the Statistics in Medicine page in the Wiley Online Library. Improving this aspect of the BART algorithm is an area of ongoing research for the authors of this paper, but at present, we are seeing sufficient performance benefits when flexibly modeling just the response surface that we deem the current algorithm worth introducing. In the meantime, we capitalize on the property highlighted in the literature on 'double robustness' [49, 50], wherein the causal estimate is correctly identified if *either* the response surface *or* the assignment mechanism is correctly specified. That is, our flexible modeling of the response surface should be sufficient for unbiased estimation even if our model for the treatment assignment is not perfect.[§]

5.5. *Posterior and Algorithm*

The BART SA algorithm is a posterior sampler for the parameters μ_{xz}, σ^2_y, β^z, and the latent variable U. A single iteration of the BART sampler produces a draw from the posterior distribution of μ_{xz} for each observation – that is, one draw of $f(X_i, Z_i)$ – and one draw of σ^2_y. It can also simultaneously produce a draw of the counterfactual, $f(X_i, 1 - Z_i)$, even if no observation was observed at this point, as it models the entirety of the response surface. Samples of the posterior of the desired treatment effect are thus made by averaging over draws of $f(X_i, 1) - f(X_i, 0)$ for appropriate subsets of the population. For example, with ATE, this involves an average over all samples, while for ATT (ATC), only the treatment (control) group is used. Regardless of the causal estimand, however, the entire sample informs the response surface fit. After the treatment effect has been estimated, draws from the posteriors of other parameters are used to update U.

This procedure is repeated until as many desired samples of the treatment effect are obtained. In practice, a number of the initial samples are discarded as 'burn-in'. Five hundred to 1000 samples are typically sufficient for burn-in, and equally as many are adequate for estimating posterior means and standard deviations.

The full model simulated by our semi-parametric SA is as follows:

$$\begin{aligned} Z \mid X, U, \beta^z &\sim \text{Bernoulli}(\Phi(X\beta^z + \zeta^z U)), \\ Y \mid U, \mu_{xz}, \sigma^2_y &\sim \text{N}\left(\mu_{xz} + \zeta^y U, \sigma^2_y\right), \\ \mu_{xz}, \sigma^2_y \mid X, Z &\sim \text{BART}(X, Z), \\ U &\sim \text{Bernoulli}(\pi^u), \\ \beta^z &\sim p(\beta^z), \end{aligned}$$

where $p(\beta^z)$ is a flat, normal, or t distribution and π^u is a hyperparameter.

An efficient method for posterior sampling in a probit regression is given by a latent variable formulation [51]. In particular, for our setting, we specify

$$\begin{aligned} Z \mid Z^* &= \text{I}_{\{Z^* \geqslant 0\}}, \\ Z^* \mid X, U, \beta^z &\sim \text{N}(X\beta^z + \zeta^z U, 1). \end{aligned}$$

Direct calculation shows that Z has the desired marginal distribution.

For brevity, we detail β^z only under a Student-t prior. A normal prior can be derived from the t by taking the limit as the degrees of freedom parameter tends to infinity. Similarly, a flat distribution results when taking the limit as the scale parameter tends to infinity. Samples are obtained for t priors by augmenting a normal prior with an unknown scale parameter:

[§] *To be clear, we are not claiming that our method is doubly robust. We are merely capitalizing on a property in causal modeling that was made more explicit in the doubly robust literature regarding the requirement to get the model right for just one of the pertinent models.*

$$\beta^z \mid \sigma^2_{\beta^z} \sim \mathrm{N}\left(0, \nu_{\beta^z}\sigma^2_{\beta^z}\Sigma_{\beta^z}\right),$$
$$\sigma^2_{\beta^z} \sim \mathrm{Inv}{-}\chi^2_{\nu_{\beta^z}}.$$

Here, ν_{β^z} and Σ_{β^z} are fixed hyperparameters and the prior obtained when marginalizing out $\sigma^2_{\beta^z}$ is multivariate t. For the most part, we use a diagonal matrix with elements consisting of the square of scale parameters, but Σ_{β^z} can be an arbitrary positive definite matrix.

We now describe how samples from the posterior of this model can be drawn. In the sequel, all distributions are conditional on X - to be concise, we omit this dependence. For any specific pair of sensitivity parameters ζ^y and ζ^z, the BART semi-parametric SA algorithm proceeds through the following steps:

(1) Run the BART sampler on $Y - \zeta^y U$ for some number of 'thinning' iterations as it updates its internal state, yielding a single sample of the vector μ_{xz} and the scalar σ^2_{y},
(2) Calculate the causal estimate using μ_{xz} and the estimated counterfactuals,
(3) Draw a sample from the conditional posterior density of the assignment mechanism parameters:

$$\begin{aligned} p\left(\beta^z, \sigma^2_{\beta^z} \mid U, Z^*\right) &\propto p(Z^* \mid U, \beta^z) p\left(\beta^z \mid \sigma^2_{\beta^z}\right) p\left(\sigma^2_{\beta^z}\right), \\ &= \exp\left\{-\frac{1}{2}\|Z^* - \zeta^z U - X\beta^z\|^2\right\}\left(\sigma^2_{\beta^z}\right)^{-p/2} \exp\left\{-\frac{1}{2\nu_{\beta^z}}\frac{1}{\sigma^2_{\beta^z}}\beta^{z\top}\Sigma^{-1}_{\beta^z}\beta\right\} \\ &\times \quad \left(\sigma^2_{\beta^z}\right)^{-(\nu_{\beta^z}/2+1)} \exp\left\{-\frac{1}{2}\frac{1}{\sigma^2_{\beta^z}}\right\}. \end{aligned}$$

In the preceding texts, p is the number of columns of X.

(a) Sample $\sigma^2_{\beta^z} \mid \beta^z \overset{\mathrm{d}}{=} \left(1 + \frac{1}{\nu_{\beta^z}}\beta^{z\top}\Sigma^{-1}_{\beta^z}\beta^z\right) \Big/ \chi^2_{\nu_{\beta^z}+p}$, where χ^2_ν is short-hand for a random variable with that distribution,
(b) Sample $\beta^z \mid U, \sigma^2_{\beta^z} \sim \mathrm{N}\left(A(Z^* - \zeta^z U), A\right), A = \left(X^\top X + \frac{1}{\nu_{\beta^z}\sigma^2_{\beta^z}}\Sigma^{-1}_{\beta^z}\right)^{-1}$.

For a normal prior, one can simply fix $\sigma^2_{\beta^z}$ to one. For a flat prior, $\Sigma^{-1}_{\beta^z}$ is the zero matrix. For stochastic EM, estimate β^z using numeric optimization.

(4) Draw independently for each observation from $U_i \mid Y, Z, \mu_{\mathrm{xz}}, \sigma^2_{\mathrm{y}}, \beta^z \sim \mathrm{Bernoulli}\left(\pi_i^{u=1} / \left(\pi_i^{u=1} + \pi_i^{u=0}\right)\right)$, where

$$\pi_i^{u=1} = \phi\left(\frac{Y_i - \zeta^y - \mu_{\mathrm{xz}}}{\sigma_{\mathrm{y}}}\right)\Phi\left(X_i\beta^z + \zeta^z\right)^{Z_i}\left[1 - \Phi\left(X_i\beta^z + \zeta^z\right)\right]^{1-Z_i}\pi^u,$$
$$\pi_i^{u=0} = \phi\left(\frac{Y_i - \mu_{\mathrm{xz}}}{\sigma_{\mathrm{y}}}\right)\Phi\left(X_i\beta^z\right)^{Z_i}\left[1 - \Phi\left(X_i\beta^z\right)\right]^{1-Z_i}(1 - \pi^u).$$

Here, ϕ is the standard normal density, and Φ is the standard normal CDF. In this step, Z_i^* has been integrated out.

(5) Draw a sample from the conditional posterior density:

$$p(Z^* \mid Z, U, \beta^z) \propto \exp\left\{-\frac{1}{2}\|Z^* - \zeta^z U - X\beta^z\|^2\right\}\prod_{i=1}^{n}\left[\mathrm{I}_{\{Z_i^* \geqslant 0, Z_i = 1\}} + \mathrm{I}_{\{Z_i^* < 0, Z_i = 0\}}\right].$$

That is, $V_i \mid Z, U, \beta^z$ are drawn independently from normal distributions, truncated above or below 0 as $Z_i = 1$ or 0, respectively.

(6) Update the BART sampler with the new $Y - \zeta^y U$.

5.6. *Simulation across different combinations of ζ^y and ζ^z*

The previous steps describe the simulation procedure for any single pair of values of ζ^y and ζ^z. More generally, we are interested in the family of posterior distributions indexed by these two parameters, approximated by pairs of values spaced on a grid. As it is reasonable to believe that small changes in

sensitivity parameters yield similar posterior distributions, the end-state for one grid cell can be used to seed the sampler in an adjacent cell. Furthermore, the algorithm as described parallelizes nicely as the grid itself is divided. This yields the global algorithm: (i) For as many units of parallelization as desired (e.g., 'cores' or processors), divide the grid into approximately equally sized and contiguous regions. (ii) Simultaneously within each region, fit the first grid cell as described earlier. And (iii) for each subsequent grid cell within each region, use the terminal state of the previous grid cell's sampler as the starting point of a new sampler. Proceed with fewer iterations of burn-in.

6. Simulation study

We conducted a large-scale simulation study to assess the ability of our method to handle particular violations of the structural and parametric assumptions necessary for causal inference. We compare the linear treatSens algorithm of [19] with our semi-parametric extension. We also compare with an approach proposed by [14], which uses a similar model to that of [19] (with the exception that a logistic link function is used instead of a probit in the assignment mechanism) but fits the model using maximum likelihood. The sensitivity parameters in that strategy serve basically the same function as our ζ_z and ζ_y but have been parameterized as partial correlations.

6.1. Simulation set-up

We divide the range of sensitivity parameters into a grid, with ζ^y ranging from 0 to 6 in increments of 0.5 and ζ^z ranging from -2.5 to 2.5 in increments of 0.25, yielding a total of 12×21 cells. Three data generating processes are used. The 'linear/linear' setting corresponds to linear specifications for both the treatment assignment mechanism and the response surface. The 'linear/nonlinear'setting corresponds to a linear specification for the treatment assignment mechanism and a nonlinear response surface. The 'nonlinear/nonlinear'setting corresponds to treatment assignment mechanism and a response surface that are both nonlinear. These models are detailed in Figure 2. We would expect to see better performance from our method in the third setting and performance similar to other methods in the first two.

Three levels of consideration for unmeasured confounding are adopted: ignoring U, estimating U in a SA, and treatment effect estimation with access to the true values of U. For all methods, the no-U fit is obtained by constraining the sensitivity parameters to 0, while in the true-U case, the parameters remain at zero but U is added as a covariate. The induced independence between response surface and assignment mechanism results in identical fits for [14] and [19] in these extreme cases. The SA results (middle panel) assess whether each sensitivity analysis algorithm can recover the true treatment effect if the correct sensitivity parameters are specified.

To further evaluate the performance of the semi-parametric treatSens algorithm, we assess two different specifications. For a fully Bayesian version, we use a t prior with three degrees of freedom, mean of 0, and a scale of 4. This distribution has both computational convenience and reasonable flexibility when fit with small sample sizes; the scale was chosen to restrict the coefficients to a range consistent with common effect sizes in probit regressions [52]. As a point of comparison with previous approaches, all of which utilize optimization in fitting the assignment mechanism, and we also include our S-EM variant. For the no-U and true-U cases, the independence between treatment and response models means that it is sufficient to run BART alone. For each of the previous grid cells and data generating models, 500

(a) Model

$$
\begin{aligned}
Y \mid X, U, Z &\sim \mathrm{N}\left(0.4x_1 + 0.4x_2 + 0.4x_3 + 0.4x_4 + z + \zeta^y u + \eta_y, 1\right), \\
Z \mid X, Y &\sim \mathrm{Bernoulli}\left(\Phi\left(-0.5 + 0.25x_1 + 0.25x_2 + 0.25x_3 + 0.25x_4 + \zeta^z u + \eta_z\right)\right), \\
U &\sim \mathrm{Bernoulli}(0.5).
\end{aligned}
$$

(b) Nonlinear terms

simulation setting	treatment (η_z)	response (η_y)
linear / linear	0	0
linear / nonlinear	0	$0.8x_1^2 + 0.8x_2^2 + 0.8x_3x_4$
nonlinear / nonlinear	$0.15x_1^2 + 0.15x_2^2 + 0.15x_3x_4$	$0.8x_1^2 + 0.8x_2^2 + 0.8x_3x_4$

Figure 2. Data generating processes and nonlinear terms for the three models used in the simulation study. All four covariates are independent standard normal random variables, and the sample size was fixed at 400.

datasets are simulated. In each dataset, each method is fit to the data, and the estimate of the treatment effect is recorded.

6.2. *Simulation results*

Figure 3 displays the simulation result. Each panel of boxplots corresponds to a combination of simulation settings reflecting the approach to confounding and the assignment mechanism/response surface combination. Each boxplot corresponds to a specific estimation approach and displays all estimates across both simulation iterations and sensitivity parameter combinations for that simulation setting.

When U is omitted from any individual analysis (left panel), the unmeasured confounder introduces a bias in proportion to the magnitude of the sensitivity parameters, ζ^y and ζ^z, and contributes to the wide spread of estimates in the left-most column. That any method shows an overall average of 1 when U is not included is an artifact of the symmetric plot range for ζ^z. None of the methods perform particularly well when U is omitted. However, semi-parametric treatSens does perform better than the linear methods

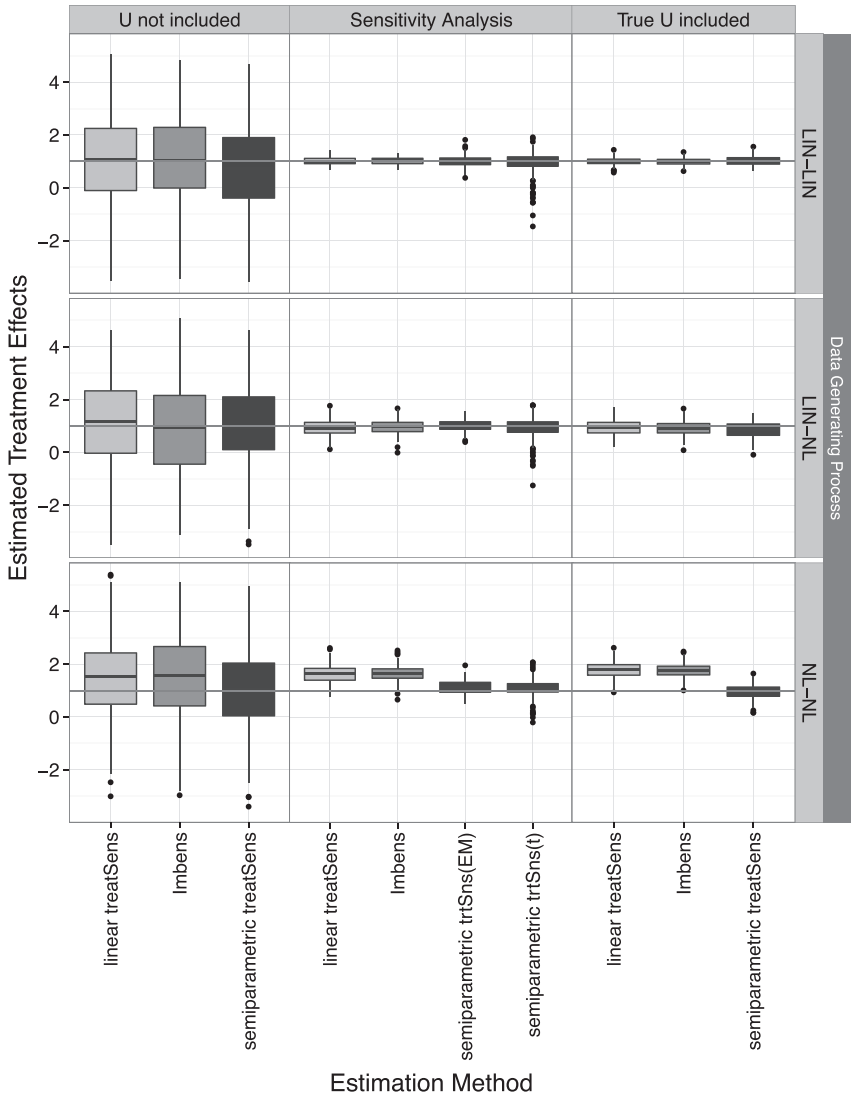

Figure 3. Box plots for the estimated treatment effects aggregated by simulations and by levels of confounding, that is, ζ^z and ζ^y grid cells. The horizontal line at 1 corresponds to the true treatment effect. From top to bottom, the rows display results from the linear/linear, linear/nonlinear, and nonlinear/nonlinear simulation settings. The left column corresponds to naïve analyses for each model that ignore U; thus, we only show results for standard Bayesian Additive Regression Trees. The middle column shows results from each sensitivity analysis approach. The right column shows results that could be achieved if U were actually observed, and thus, we only have one Bayesian Additive Regression Trees fit again.

of [19] and [14] in the nonlinear/nonlinear simulation scenario. This is due to BART's ability to flexibly model arbitrary nonlinear response surfaces.

When performing a full SA (middle panel), we observe a vast reduction in the variability of estimates (relative to the setting that ignores U) because of a decrease in omitted-confounder bias. The comparable performance across methods in linear/linear and linear/nonlinear settings demonstrates that the correct treatment effect can be obtained when at least one of the treatment or response models is correct. Said another way, when the response model is nonlinear but the assignment mechanism is linear, the nonlinear terms in the response surface do not act as confounders and thus can be ignored without introducing bias.

In the setting with nonlinear treatment assignment and response surface, semi-parametric treatSens performs much better than its competitors. The fully Bayesian implementation of the semi-parametric treatSens still exhibits slightly greater variability than the other methods; however, the results are centered around the true treatment effect estimate. The S-EM variant of this algorithm exhibits less variability than the linear methods except in the linear/linear scenario.

Finally, the minimal change in results when U is directly included in the fit (right panel) demonstrates that these methods effectively recover the effect of the unmeasured confounder. Note again, however, that even with U included, the other methods fail in the nonlinear–nonlinear setting because neither the treatment assignment nor response surface is modeled correctly. Overall Figure 3 suggests that a semi-parametric approach to sensitivity analysis can be crucially important in the presence of nonlinear confounding.

Figure 3 aggregates simulation results across levels of the sensitivity parameters. In contrast, Figure 4 disaggregates the results by combinations of sensitivity parameters and displays them in the form of a heat map. The closer the color is to blue in a given grid square, the larger the treatment effect estimate; the closer to red, the smaller (greater in negative value) the estimate. Lack of color indicates treatment effect

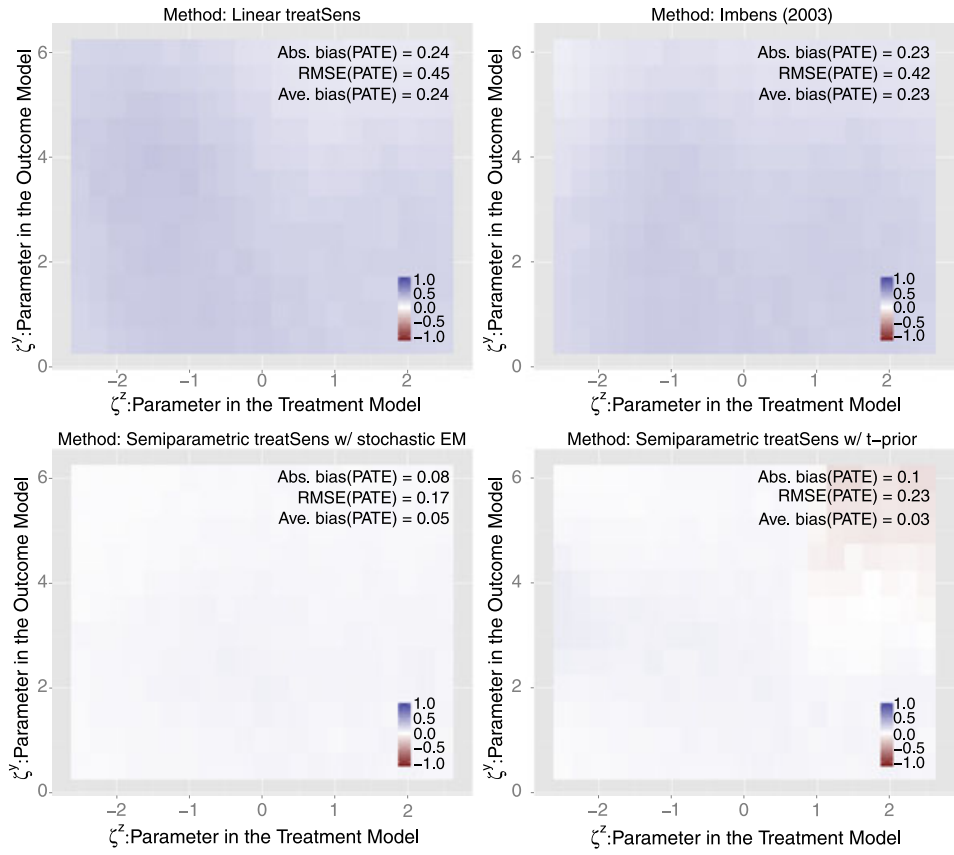

Figure 4. Heat maps of the bias in the estimated treatment effect for all four sensitivity analysis techniques in the case where the treatment and response models are nonlinear. Each cell is the average of 500 simulations with the level of unmeasured confounding given by the x and y axes, expressed in units of the standard deviation of the response variable. Reported biases are averages across all grid cells. 'Abs. bias' is calculated by taking absolute values first, so that overestimation in one region is not offset by underestimation in another.

estimates close to zero. Given the increased complexity of these plots, we restrict focus to a comparison of our four SA methods in the nonlinear–nonlinear setting, as in the others, there was little variation across combinations of the sensitivity parameters.

This figure reveals that both the linear methods yield positively biased treatment effect estimates for most combinations of sensitivity parameters. The plot corresponding to the stochastic-EM implementation of our semi-parametric SA approach, on the other hand, demonstrates little to no bias across the board.

The fully Bayesian semi-parametric SA is a more interesting case with slight negative bias when both sensitivity parameters are large and positive. This occurs because in this region, overlap across treatment groups (with respect to U) is compromised and the response surface for the unsupported group is regressed to its prior mean. The fact that this bias is not present in the S-EM version of our algorithm suggests that the estimate is (in part) a by-product of uncertainty in the treatment mechanism. Indeed, for extreme levels of both ζ^y and ζ^z, the linear methods failed to converge in the linear/linear and the nonlinear/linear model (this occurred in less than 0.8% of cells, which were simply omitted from analysis). When lack of overlap is sufficiently pronounced, not only will the estimates be biased but also the treatment model may become separable, with some subset of the covariates perfectly able to predict inclusion in treatment or control. One implication of this is that researchers need to pay close attention to the overlap across groups and may want to additionally implement methods such as those suggested in [28].

7. Application: Effectiveness of diuretics on high blood pressure using Third National Health and Nutrition Examination Survey

Now, we investigate how our semi-parametric SA framework works on real-world data. Specifically, we examine the effectiveness of anti-hypertensive drugs on the level of blood pressure using data from the NHANES III [11].

7.1. Background

High blood pressure (HBP) is one of the most common and most lethal diseases in the USA. In 2006, about one third of US adults were affected by HBP [53], and HBP is known to be a primary risk factor for life-threatening cardiovascular diseases such as heart failure, coronary heart disease, and stroke [54]. Strikingly, HBP accounted for 17.8% of US deaths in 2006, a rate which represents a 19.5% increase from 1996 [53]. The high prevalence of HBP and high fatality rate of HBP-related diseases make the development of anti-hypertensives, one of the most lucrative businesses in the drug industry. As of 2010, the market for anti-hypertensives amounted to about $27bn, and 67 commonly used anti-hypertensive drugs are available as of this writing [53].

Every anti-hypertensive drug that is sold in the USA must pass the US Federal Drug Administration's drug review process. Although the multi-phase trial process of Federal Drug Administration's drug approval provides considerable information about the efficacy of the drug under idealized circumstances, voluntary trial subjects are not necessarily representative of the population of people suffering from HBP, and so, the effect of treatment in the general population may differ from that observed in a clinical trial. Moreover, continued monitoring with observational studies yields important information about the effectiveness of the drug given real-world prescription and adherence patterns.

Risk factors for HBP have been an active area of research in fields as varied as anthropology and sociology. A multitude have been identified, including socioeconomic factors such as age, gender, education, and income; lifestyle factors such as high sodium intake, low potassium intake, obesity, and modernization; daily stressors like discrimination and racism; insufficient coping mechanisms such as a lack of kin support, loss of traditional culture, and low education; and finally hereditary factors such as family history or genetics. The American Heart Association provides a recent overview in [55]. This body of research informs our choice of confounding covariates.

Many studies have identified nonlinear relationships and interdependencies among risk factors and HBP [56–59]. For example, income has a nonlinear relationship with blood pressure, and different nonlinear relationships exist for each sex [60]. Simple linear regression with a linear combination of covariates, ignoring both nonlinear terms and interactions, may fail to control for aspects of the response surface. Moreover, identification of these features may be difficult in high-dimensional space. Even if we detect from residual plots that there is a problem, the solution – for instance, adding nonlinear terms or interactions – may be elusive. In our SA framework, mis-estimation of the response surface could lead to

miscalculation of the probabilities from which the unobserved confounders are drawn and thus biased estimates of the treatment effect corresponding to any given set of sensitivity parameters. BART has been shown to be effective in flexibly modeling complicated, nonlinear response surfaces. Consequently, comparing SA results arising from assuming a simple linear model to the BART counterpart can demonstrate the importance of identifying the correct response surface.

7.2. Data and variables

Our data come from the NHANES III, one of the most extensive surveys on Americans' health and nutritional status and a source of numerous important findings [61–63]. In NHANES III, socioeconomic background, medical record, dietary pattern, daily activities, and other health-related issues are recorded for respondents of age 2 months or older. A 4-hour health exam is also performed in mobile examination centers. In order to restrict the source of bias to misspecification of the parametric model, we minimize heterogeneities between treatment groups by excluding healthy individuals. This is accomplished by defining the control group as those who reported being informed by a doctor that they have HBP but were not taking any medicine. Because congenital heart problems represent a distinct test case, we further limit the data to adults at least 17 years old.

Using these data, we highlight two illustrative cases: one in which the treatment effect is estimated as non-significant with a linear model while it is estimated as significant and negative with BART, and another which the two have an opposite relationship. Accordingly, we selected two sets of treatment and dependent variables. We first present the results of the effect of 'taking two or more anti-hypertensives' on average diastolic blood pressure. Then, we present the results of the effect of 'taking beta blocker and diuretics' on average systolic blood pressure. Both treatments are defined from reported prescription drug use, and those drugs primary use classes. Because the American Heart Association suggests that doctors prescribe an anti-hypertensive with thiazide-type diuretics for those who cannot lower their blood pressure by modifying their lifestyle [54], the patients in these examples can be thought of as following a standard regimen.

Based on the findings from previous studies discussed earlier, we include the following pre-treatment variables as covariates: An indicator variable for whether the respondent is female, an indicator for whether the respondent is non-Hispanic white, an indicator for whether she or he is black, an indicator for whether she or he is Hispanic, age (in months), household size, number of years of education completed, indicators for whether she or he is married, whether s/he is widowed, whether s/he is separated (using never married as a baseline category), logged annual household income, pack years (number of packs smoked everyday multiplied by number of years smoked), body mass index((mass (lb)/height $(\text{in})^2$) $\times$ 703), radial pulse rate (beats/min), sodium intake (mg), potassium intake (mg), sodium–potassium ratio, alcohol intake (g), an indicator for whether she or he has health insurance, and finally the frequency of meeting with friends or relatives per year. Observations incomplete with respect to the outcome or the control variables were excluded from analysis.

7.3. Results

Figure 5 shows the SA for the effect of taking two or more anti-hypertensives on diastolic blood pressure, while Figure 6 shows the analysis for the effect of taking beta blockers and diuretics on systolic blood pressure. The results in the left panel are obtained from the SA in which covariate effects are linear and additive, as detailed by [19]. The right panel shows the results of our semi-parametric SA that accounts for nonlinear effects and the interaction effects of the covariates on the response surface. For all results in this section, we use a 10 by 20 grid with 5000 draws of the unmeasured confounder, U, per cell.

The various types of contour lines show the estimated treatment effect for the levels of confounding under the sensitivity parameters values on the relevant axes. The thin black lines report the basic effect, while the colored lines highlight specific levels of interest. Specifically, the blue lines labeled with 'N.S.' demarcate the point at which the estimated effect is no longer statistically significant at the 5% level. These bracket a red line, which shows the confounding necessary to drive the estimate to 0. Finally, the thick gray line corresponds to the treatment effect estimate that would arise with an unmeasured confounder whose strength is equivalent to that of the covariates whose marginal effect sizes are of the greatest magnitude. The naïve treatment effect estimate is reported next to the horizontal line at the base of the y axis, in the lower right of any plot.

Symbols in these figures correspond to the marginal effect sizes of covariates from naïve analyses for the response model (y axis) and treatment model (x axis). For any parametric fit, including both levels

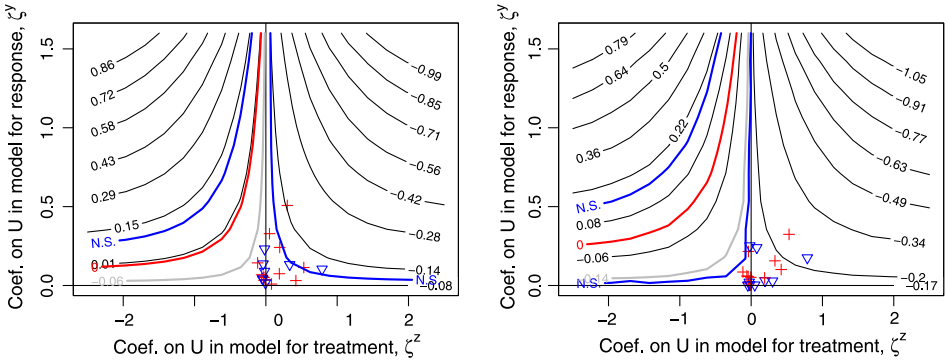

Figure 5. Sensitivity analysis results for the effects of taking two or more anti-hypertensives on diastolic blood pressure using the Third National Health and Nutrition Examination Survey data. The left panel displays the results of fully parametric sensitivity analysis in which the response surface is fitted with a linear combination of covariates. The right panel displays the results of the semi-parametric sensitivity analysis.

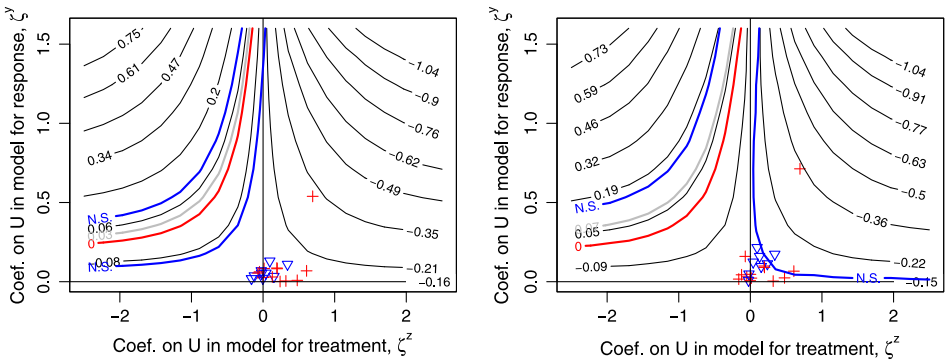

Figure 6. Sensitivity analysis results for the effects of taking beta blockers and diuretics on systolic blood pressure using the Third National Health and Nutrition Examination Survey data. The left panel displays the results of the fully parametric sensitivity analysis in which the response surface is fitted with a linear combination of covariates. The right panel displays the results of the semi-parametric sensitivity analysis.

of the linear SA and the assignment mechanism in the semi-parametric model, the marginal effects are simply the coefficients in a regression. For the BART fit of the response surface, marginal effects are obtained by estimating the average treatment effect when a covariates goes from −0.5 standard deviations below average to 0.5 above. The observed covariates that have a negative association with blood pressure have been rescaled (multiplied by −1) so that the estimated coefficients will be positive; these are represented by ▽ on the plot. Finally, the coefficients of all continuous covariates are standardized to facilitate comparisons with the hypothesized unmeasured confounder.

The results of the linear SA in the left panel of Figure 5 indicate that, absent unmeasured confounding, the treatment effect estimate would be about −0.08, that is, close to zero. Moreover, the sensitivity parameters for an unobserved binary confounder only need to be stronger than the coefficient of age (the strongest observed confounder plotted in the upper right) to reduce the treatment effect to zero. On the other hand, the results of the semi-parametric SA in the right panel show the statistically significant and negative treatment effect of −0.17 if the ignorability assumption holds. Although these naïve treatment effect estimates are somewhat different, both results are fairly sensitive to the effect of an unobserved confounder. For instance, a confounder with a coefficient in the treatment model of −1 and coefficient of 0.5 for the outcome model would change the signs of the negative treatment effect estimates in the both panels.

Figure 6 shows the results of the SA for the effect of taking beta blockers and diuretics on systolic BP. While the linear SA and the semi-parametric SA produce similar naïve treatment effect estimates (−0.16 and −0.15, respectively), the results of the latter are more sensitive to unobserved confounding than the former. For instance, a confounder with a coefficient in the treatment model of −0.5 and coefficient of 0.25 for the outcome model would not change the statistical significance of the results using the linear

SA, while using a semi-parametric SA, the naïve treatment effect of −0.15 is already not statistically significant.

These two figures stress that our inference on the sensitivity of the treatment effects to unmeasured confounding can substantively change, depending on how the response surface is predicted. Forcing the use of a linear response surface can induce undue confidence in the results, as in Figure 6 or increase the sensitivity of estimated treatment effects that are captured more robustly when the response surface is allowed to be nonlinear, as in Figure 5. The direction and magnitude of the bias introduced by misspecification of the response surface will determine whether the semi-parametric approach will increase or decrease apparent sensitivity to unmeasured confounding.

8. Conclusion

More often than not, it is impractical to implement randomized experiments to address many of the most interesting causal questions. The alternative approach of using observational studies to draw causal conclusions requires structural as well as functional assumptions. These structural assumptions are typically not trivially plausible, which motivates analysis of the sensitivity of causal estimates drawn from observational studies to violations of these assumptions, in particular of ignorability. In order for such strategies to yield results that are useful to applied researchers, these SAs should be easily interpretable, preferably employing sensitivity parameters whose magnitudes are calibrated based on contextual information (for instance, analogous parameter estimates for observed covariates). However, these goals can be more difficult to achieve if one is forced to rely on parametric models, as the potential for model misspecification introduces its own biases. We sidestep this issue by allowing for a nonparametric fit of the relationship between the outcome and the observed covariates via the BART algorithm. This approach appears to be competitive with existing approaches when no nonlinear confounding exists and to outperform these approaches in the presence of nonlinear confounding. Moreover, the procedure has been integrated into the treatSens package for the R programming language available, on the Comprehensive R Archive Network.

Acknowledgements

This research was partially supported by Institute of Education Sciences grants R305D110037 and R305B120017 and JSPS KAKENHI Grant Number 15K16977.

References

1. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-experimental Designs*. Houghton Mifflin Company: Boston, MA, 2002.
2. Angrist JD, Pischke JS. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press: Princeton, NJ, 2008.
3. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 1996; **91**:444–472.
4. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: New York, 2007.
5. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* 1959; **22**:173–203.
6. Bross ID. Spurious effects from an extraneous variable. *Journal of Chronic Diseases* 1966; **19**(6):637–647.
7. Bross ID. Pertinency of an extraneous variable. *Journal of Chronic Diseases* 1967; **20**(7):487–495.
8. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)* 1983a; **45**(2):212–218.
9. Manski C. Nonparametric bounds on treatment effects. *American Economic Review Papers and Proceedings* 1990; **80**: 319–323.
10. Rosenbaum PR. *Observational Studies*. Springer: New York, 2002.
11. Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data III, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Hyattsville, MD, 1997. Available from: http://www.cdc.gov/nchs/nhanes/nh3data.htm [Accessed on 5 March 2014].
12. Rosenbaum PR. Covariance adjustment in randomized experiments and observational studies. *Statistical Science* 2002; **17**(3):286–327.
13. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* 1987b; **74**(1):13–26.

14. Imbens G. Sensitivity to exogeneity assumptions in program evaluation. In *The American Economic Review: Papers and Proceedings of the One Hundred Fifteenth Annual Meeting of the American Economic Association*, Vol. 93: New York, NY, 2003; 126–132.
15. Gastwirth JL, Krieger AM, Rosenbaum PR. Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* 1998; **85**(4):907–920.
16. Greenland S. Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology* 1996; **25**(6): 1107–1116.
17. Rosenbaum PR. Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association* 2010; **105**:692–702.
18. Harada M. Generalized sensitivity analysis. Technical Report, New York University, New York, NY, 2013.
19. Carnegie NB, Harada M, Hill J. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness* 2016; **9**(3):395–420.
20. Rosenbaum PR, Silber JH. Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association* 2009; **104**:1398–1405.
21. Ichino A, Mealli F, Nannicini T. From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics* 2008; **23**(3):305–327.
22. Ho DK, Imai K, King G, Stuart E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007; **15**(3):199–236.
23. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 2003; **71**:1161–89.
24. Hill JL, Weiss C, Zhai F. Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research* 2011; **46**:477–513.
25. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research* 2015. DOI:10.1177/0962280215584401. Available from: http://smm.sagepub.com/content/early/2015/04/30/0962280215584401.abstract.
26. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 1998; **66**:315–322.
27. Hill J. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 2011; **20**(1):217–240.
28. Hill J, Su YS. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on childrens cognitive outcomes. *Annals of Applied Statistics* 2013; **7**(3):1386–1420.
29. Karabatsos G, Walker SG. A Bayesian nonparametric causal model. *Journal of Statistical Planning and Inference* 2012; **142**(4):925–934.
30. Rubin DB. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine* 2009; **28**(9):1420–1423.
31. Pearl J. On a class of bias-amplifying variables that endanger effect estimates. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, Catalina Island, CA, 2010, 425–432. Available from: http://ftp.cs.ucla.edu/pub/stat_ser/r356.pdf [Accessed on 2 February 2016].
32. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Strmer T. Variable selection for propensity score models. *American Journal of Epidemiology* 2006; **163**:1149–1156.
33. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.)* 2009; **20**(4):512–522.
34. Brooks JM, Ohsfeldt RL. Squeezing the balloon: propensity scores and unmeasured covariate balance. *Health Services Research* 2013; **48**(4):1487–1507.
35. Carnegie NB, Harada M, Dorie V, Hill J. *treatsens: Sensitivity Analysis for Causal Inference*, 2015. Available from: http://CRAN.R-project.org/package=treatSens [Accessed on 14 July 2015], R package version 2.0.
36. R Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. Available from: http://www.R-project.org/ [Accessed on 14 July 2015], ISBN 3-900051-07-0.
37. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Annals of Applied Statistics* 2010; **4**(1):266–298.
38. Rubin DB. Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics* 1978; **6**:34–58.
39. Barnow BS, Cain GG, Goldberger AS. Issues in the analysis of selectivity bias. In *Evaluation Studies*, Stromsdorfer E, Farkas G (eds), Vol. 5. Sage: San Francisco, 1980; 42–59.
40. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* 1986; **15**(3):413–419.
41. Lechner M. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies*, Lechner M, Pfeiffer F (eds), ZEW Economic Studies, vol. 13. Physica-Verlag: HD, 2001; 43–58.
42. Green DP, Kern HL. Modeling heterogeneous treatment effects in survey experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly* 2012; **76**(3):491–511.
43. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*, Vol. 43. Chapman and Hall/CRC: Boca Raton, FL, 1990.
44. Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning*. MIT Press: Cambridge, MA, 2006.
45. Wand MP, Jones MC. *Kernel Smoothing*, Vol. 60. Chapman and Hall/CRC: Boca Raton, FL, 1994.
46. Chipman H, McCulloch R. *BayesTree: Bayesian methods for Tree Based Models*, 2010. Available from: http://CRAN.R-project.org/package=BayesTree [Accessed on 3 November 2014], R package version 0.3-1.1.
47. Dorie V, Chipman H, McCulloch R. *DBARTS: Discerete Bayesian Additive Regression Trees Sampler*, 2014. Available from: http://CRAN.R-project.org/package=dbarts [Accessed on 13 November 2014], R package version 0.8-5.

48. Celeux G, Diebolt J. The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational Statistics Quarterly* 1985; **2**(1):73–82.
49. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 1995; **90**:122–129.
50. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science* 2007; **22**:523–580.
51. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* 1993; **88**(422):669–679.
52. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2008; **2**(4):1360–1383.
53. Lloyd-Jones D, Adams RJ, Brown TM, Carnethon M, Dai S, De Simone G, Ferguson T, Ford E, Furie K, Gillespie C, Go A, Greenlund K, Haase N, Hailpern S, Ho PM, Howard V, Kissela B, Kittner S, Lackland D, Lisabeth L, Marelli A, McDermott MM, Meigs J, Mozaffarian D, Mussolino M, Nichol G, Roger VL, Rosamond W, Sacco R, Sorlie P, Stafford R, Thom T, Wasserthiel-Smoller S, Wong ND, Wylie-Rosett J, on behalf of the American Heart Association Statistics Committee, Stroke Statistics Subcommittee. Heart disease and stroke statistics-2010 update: A report from the american heart association. *Circulation* 2010; **121**(7):e46–e215.
54. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL, Jr., Jones DW, Materson BJ, Oparil S, Wright JT, Jr., Roccella EJ, the National High Blood Pressure ducation Program Coordinating Committee. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: the JNC 7 report. *Journal of the American Medical Association* 2003; **289**(19):2560–2571.
55. Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Borden WB, Bravata DM, Dai S, Ford ES, Fox CS, Franco S, Fullerton HJ, Gillespie C, Hailpern SM, Heit JA, Howard VJ, Huffman MD, Kissela BM, Kittner SJ, Lackland DT, Lichtman JH, Lisabeth LD, Magid D, Marcus GM, Marelli A, Matchar DB, McGuire DK, Mohler ER, Moy CS, Mussolino ME, Nichol G, Paynter NP, Schreiner PJ, Sorlie PD, Stein J, Turan TN, Virani SS, Wong ND, Woo D, Turner MB. Heart disease and stroke statistics-2013 update: A report from the american heart association. *Circulation* 2013; **127**(1):e6–e245.
56. Angeli F, Reboldi G, Verdecchia P. From Apennines to Andes: does body mass index affect the relationship between age and blood pressure? *Hypertension* 2012; **60**(1):6–7.
57. Fillenbaum GG, Blay SL, Pieper CF, King KE, Andreoli SB, Gastal FL. The association of health and income in the elderly: experience from a southern state of brazil. *PloS one* 2013; **8**(9):e73930.
58. Gurven M, Blackwell AD, Rodríguez DE, Stieglitz J, Kaplan H. Does blood pressure inevitably rise with age? Longitudinal evidence among forager-horticulturalists. *Hypertension* 2012; **60**(1):25–33.
59. Zhang Y, Li H, Liu Sj, Fu Gj, Zhao Y, Xie YJ, Zhang Y, Wang Yx. The associations of high birth weight with blood pressure and hypertension in later life: a systematic review and meta-analysis. *Hypertension Research* 2013; **36**(8):725–735.
60. Rehkopf DH, Krieger N, Coull B, Berkman LF. Biologic risk markers for coronary heart disease: nonlinear associations with income. *Epidemiology* 2010; **21**(1):38–46.
61. Alexander CM, Landsman PB, Teutsch SM, Haffner SM. NCEP-defined metabolic syndrome, diabetes, and prevalence of coronary heart disease among NHANES III participants age 50 years and older. *Diabetes* 2003; **52**(5):1210–1214.
62. Flegal KM, Carroll MD, Kuczmarski RJ, Johnson CL. Overweight and obesity in the united states: prevalence and trends, 1960-1994. *International Journal of Obesity and Related Metabolic Disorders: Journal of the International Association for the Study of Obesity* 1998; **22**(1):39–47.
63. Hollowell JG, Staehling NW, Flanders WD, Hannon WH, Gunter EW, Spencer CA, Braverman LE. Serum TSH, T4, and thyroid antibodies in the United States population (1988 to 1994): National Health and Nutrition Examination Survey (NHANES III). *The Journal of Clinical Endocrinology & Metabolism* 2002; **87**(2):489–499.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site.

