Matching on Generalized Propensity Scores with Continuous Exposures

Xiao Wu
Department of Biostatistics, Harvard T.H. Chan School of Public Health
and
Fabrizia Mealli
Department of Statistics, Informatics, Applications, University of Florence
and
Marianthi-Anna Kioumourtzoglou
Department of Environmental Health Sciences,
Mailman School of Public Health, Columbia University
and
Francesca Dominici, Danielle Braun
Department of Biostatistics, Harvard T.H. Chan School of Public Health

February 7, 2019

Abstract

Generalized propensity scores are commonly used to adjust for confounding when estimating the causal effects of a continuous treatment (or exposure) in observational studies. Existing approaches include using the estimated generalized propensity score a) as a covariate in the outcome model, b) for inverse probability of treatment weighting, or c) in doubly robust estimators. However, these approaches have the following limitations. First, they require that either the generalized propensity score model or the outcome model, or both, are correctly specified. Second, both inverse probability of treatment weighting and doubly robust approaches rely on weighting and are, therefore, sensitive to extreme values of the estimated generalized propensity score. Third, assessing covariate balance when using these approaches is not straightforward. In this paper we propose an innovative caliper matching approach that uses the generalized propensity score in settings with continuous exposures. We first introduce an assumption of identifiability called local weak unconfoundedness that is less stringent than what is currently proposed in the literature. Under this assumption and under mild smoothness conditions we provide theoretical results that guarantee that our

proposed matching estimators attain consistency and asymptotic normality. Importantly, we adapt two measures of covariate balance into the continuous matching framework. In simulation studies, our proposed matching estimator outperforms existing methods under settings of model misspecification and/or in presence of extreme values of the estimated generalized propensity score in terms of bias reduction and root mean squared error, and overall improves the covariate balance. We apply the proposed method to Medicare Part A data in New England to estimate the causal effect of long-term exposure to fine particles ($PM_{2.5}$), a continuous exposure, on mortality for the period from 2000 to 2012.

Keywords: Causal inference, Exposure-response function, Generalized propensity score, Matching, Observational studies

1 Introduction

Long-term exposure to air pollution has been associated with adverse health outcomes in many observational studies (Beelen et al. 2014, Kioumourtzoglou et al. 2016, Di et al. 2017), yet limited literature is available to estimate the effects of air pollution exposure in a causal framework (Wang et al. 2016, 2017). Estimating the causal effects of air pollution exposure on health outcomes is challenging as 1) there are a large set of covariates that are associated with both the exposure and outcome of interest (potential confounders), 2) the exposure is continuous and flexible estimation of an exposure response curve is highly desirable.

When trying to estimate effects in observational studies, confounding occurs due to lack of randomization. Failure to properly account for confounders in the analysis may lead to substantial bias in effect estimates. Although most studies adjust for confounding, many do so by fitting a regression model relating the outcome to the exposures and covariates. These standard regression methods mix the design and analysis stages, leading more likely to deviations from causality (Rubin et al. 2008). In many observational studies, the treatment (or exposure) is continuous in nature. Estimating the causal effects in these studies is challenging as 1) there is a large set of covariates that are associated with both the exposure and outcome of interest (potential confounders), and 2) one has to allow for flexible estimation of the exposure-response function on a continuous scale.

In practice, most studies adjust for confounding by fitting a regression model relating the outcome to the exposures and covariates. These standard regression methods mix the design and analysis stages, more likely leading to deviations from causality (Rubin et al. 2008). Under the potential outcome framework for causal inference, the design stage (where the goal is to adjust for confounding by achieving covariate balance) and the analysis stage (where we aim at estimating causal effects) are distinct (Imbens & Rubin 2015). A common approach for confounding adjustment in the design stage is using propensity scores; the probability of a unit being assigned to a particular treatment (or exposure in this setting), given the pre-treatment covariates. Using propensity scores to adjust for confounding was first introduced by Rosenbaum & Rubin (1983). After this seminal paper, advanced propensity score techniques have been developed to estimate causal effects in observational studies. These are reviewed by, for example, Harder et al. (2010). The main limitation of these approaches is that they were developed for binary exposures. There

are settings in which there is interest in estimating exposure effects of categorical exposures. To handle categorical exposures, Imbens (2000) developed the generalized propensity score, a natural analogue to propensity scores which uses multinomial regression models instead of binary regression models to model the distribution of the exposure given covariates, and showed that it can be used for inverse probability of treatment weighting. Although there is no natural analogue for matching and subclassification for categorical generalized propensity score (Rassen et al. 2013), Yang et al. (2016) propose an alternative way to estimate causal effects in these settings.

Hirano & Imbens (2004) propose an extension of the categorical generalized propensity score to continuous exposures: under correct specification of both generalized propensity score and outcome models, adjustment for confounding is obtained by including the estimated generalized propensity score as a covariate in the outcome model. Robins et al. (2000) propose an approach using marginal structure models in which exposure effects are estimated using a class of inverse probability of treatment weighting estimators. However, this approach also requires the correct specification of both generalized propensity score and outcome models. A doubly robust estimator, first proposed by Robins et al. (1994), is a class of augmented inverse probability of treatment weighting estimator that is more robust to model misspecification, since the parameters can be consistently estimated even when only one of the two models is correctly specified (Cao et al. 2009). However, the standard doubly robust approach for exposure-response estimation relies on parametric model specifications, and performs poorly when both the (generalized) propensity score model and the outcome model are misspecified, which is likely to happen in applications (Kang et al. 2007, Waernbaum 2012). Kennedy et al. (2017) recently proposed non-parametric approaches for doubly robust estimation of continuous exposure effects, which aim at reducing model dependency.

The matching estimator, another popular estimation technique in the binary/categorical exposure settings (Rosenbaum & Rubin 1983, Yang et al. 2016), has the following properties: 1) it is robust to misspecifications of the generalized propensity score model, especially in the presence of extreme values of the generalized propensity score (Waernbaum 2012), 2) it does not have any dependence on the outcome model specification, and 3) it allows for the straightforward assessment of covariate balance. Abadie & Imbens (2006) show that the matching estimator is consistent and asymptotically normal when matching on a scalar function, such as the (generalized) propensity

score. However, such matching estimator has never been extended and implemented for continuous exposures. Extending the matching estimator to a continuous exposure is not straightforward as matching observations with the exact exposure level is difficult (Flores et al. 2007). In our work, we focus on settings for which we have a continuous exposure, and propose a novel generalized propensity score caliper matching framework that jointly matches on both the estimated generalized propensity score and exposure levels to adjust for confounding. Our proposed approach aims at reducing model dependence in both the design and analysis stages.

2 The Generalized Propensity Score Function

Let N denote the study sample size. For each unit $j \in \{1, \ldots, N\}$, let $\mathbf{C}_j$ denote the pretreatment covariates for unit j, which is characterized by a M-vector $(C_{1j}, \ldots, C_{Mj})$; W_j denote the continuous exposure for unit j, $W_j \in \mathbb{W}$ with a range $[w^0, w^1]$; $Y_j(w)$ denote the counterfactual outcome for unit j at the exposure level w; and $p_j\{W \mid \mathbf{C}, Y(w)\}$, for all $w \in \mathbb{W}$, denote the assignment mechanism defined as the conditional probability density of each exposure level given the covariates and potential outcomes. One target estimand is the population average causal exposure-response function defined on the specific range of the exposure levels $w \in [w^0, w^1]$, $\mu(w) = E\{Y_j(w)\}$. Under the potential outcomes framework (Rubin 1974) which was adapted to continuous exposures (Hirano & Imbens 2004, Kennedy et al. 2017), we establish the following assumptions of identifiability:

Assumption 1 (Consistency) *$W = w$ implies $Y = Y(w)$.*

Assumption 2 (Overlap) *For all values of $\mathbf{c}$, the density function of receiving any possible exposure $w \in \mathbb{W} = [w^0, w^1]$ is positive: $f(w \mid \mathbf{c}) > 0$ for all w, $\mathbf{c}$.*

This assumption guarantees that for all possible values of pretreatment covariates, $\mathbf{c}$, $\mu(w)$ can be estimated for each exposure w without relying on extrapolation.

Condition 1 (Weak Unconfoundedness) *The assignment mechanism is weakly unconfounded if for all $w \in \mathbb{W}$, in which w is continuously distributed with respect to the Lebesgue measure on $\mathbb{W} = [w^0, w^1]$; $W_j \perp\!\!\!\perp Y_j(w) \mid \mathbf{C}_j$.*

Condition 1 implies that we do not require (conditional) independence of potential outcomes $Y_j(w)$ for all $w \in [w^0, w^1]$, jointly, that is $W_j \perp\!\!\!\perp \{Y_j(w)\}_{w\in[w^0,w^1]} \mid \mathbf{C}_j$. Instead, we only require conditional independence of the potential outcome $Y_j(w)$ for a given exposure level w. Most causal inference studies using continuous exposures rely on this condition (Robins et al. 2000, Hirano & Imbens 2004, Imai & Van Dyk 2004, Flores et al. 2007, Galvao & Wang 2015, Kennedy et al. 2017).

We introduce Assumption 3 called *Local Weak Unconfoundedness* which is less stringent than the weak unconfoundedness assumption.

Assumption 3 (Local Weak Unconfoundedness) *Let $I_j(.)$ be an indicator variable indicating if exposure level $W_j = \tilde{w}$ or not for $\tilde{w} \in [w - \delta, w + \delta]$, where δ is the caliper defined as the radius of the neighborhood around w, and follows a positive sequence tending to zero as $N \to \infty$. The assignment mechanism is locally weakly unconfounded if for all $w \in \mathbb{W}$, for which w is continuously distributed with respect to the Lebesgue measure on $\mathbb{W} = [w^0, w^1]$, then $\{I_j(\tilde{w})\}_{\tilde{w}\in[w-\delta,w+\delta]} \perp\!\!\!\perp Y_j(w) \mid \mathbf{C}_j$.*

The local refers to the fact that we define an exposure set $\tilde{w} \in [w - \delta, w + \delta]$ that contains a neighborhood around w. This assumption is weaker than Condition 1, and can be deduced from Condition 1 as $\{I_j(\tilde{w})\}_{\tilde{w}\in[w-\delta,w+\delta]}$ is measurable with respect to the σ-algebra generated by W_j. It is natural to couple this assumption with the following smoothness assumption.

Assumption 4 (Smoothness) *Suppose the average exposure-response function $E\{Y_j(w)\}$ is continuous with respect to w, and $h \geq \delta$, where h is a sequence tending to zero as $N \to \infty$ and δ as previously defined, then $\lim_{h\to 0} E\{Y_j(w-h)\} = \lim_{h\to 0} E\{Y_j(w+h)\} = E\{Y_j(w)\}$.*

We follow the generalization of the propensity score from binary exposure to continuous exposure as proposed by Hirano & Imbens (2004).

Definition 1 *The generalized propensity score is the conditional density function of the exposure given pretreatment covariates: $\mathbf{e}(\mathbf{c}_j) = \{f_{W|\mathbf{C}_j}(w \mid \mathbf{c}_j), \forall w \in [w^0, w^1]\}$. The individual $e(w, \mathbf{c}_j) = f_{W|\mathbf{C}_j}(w \mid \mathbf{c}_j)$ are called realizations of $\mathbf{e}(\mathbf{c}_j)$.*

The following Lemmas show that 1) local weak unconfoundedness holds when we condition on the generalized propensity score. 2) The population average causal exposure-response function, that is our target estimand, is identifiable under Assumptions 1-4.

Lemma 1 (Local Weak Unconfoundedness Given Generalized Propensity Score) *Suppose the assignment mechanism is locally weakly unconfounded. Then for all* $w \in \mathbb{W} = [w^0, w^1]$ *and* $\tilde{w} \in [w - \delta, w + \delta]$, $I_j(\tilde{w}) \perp\!\!\!\perp Y_j(w) \mid e(\tilde{w}, \mathbf{C}_j)$.

Lemma 2 (Average Causal Exposure-response Function) *Suppose the assignment mechanism is locally weakly unconfounded. Then for all* $w \in \mathbb{W} = [w^0, w^1]$, $\mu(w) = E\{Y_j(w)\} = \lim_{\delta \to 0} E[E\{Y_j^{obs} \mid e(w_j, \mathbf{C}_j), w_j \in [w - \delta, w + \delta]\}]$.

Lemma 2 allows us to estimate the exposure-response function at the exposure level w, $E\{Y_j(w)\}$ as an average of conditional expectations for the specific value of the generalized propensity score $e(w_j, \mathbf{C}_j)$ and exposure w_j in an exposure set $[w - \delta, w + \delta]$. By conditioning on the single value of the generalized propensity score $e(w_j, \mathbf{C}_j)$, the population average causal exposure-response function is still identifiable under the local weak unconfoundedness assumption (Yang et al. 2016). Proofs of both Lemmas are presented in the Supplementary Material.

3 Matching Framework

3.1 General Matching Function

The ultimate objective for matching is to construct matched datasets that mimic a randomized experiment as closely as possible by achieving good covariate balance. In the categorical setting, Yang et al. (2016) propose a generalized propensity score matching approach which creates matched datasets consisting of replicated units representing the quasi-experimental arm for each exposure category. In the continuous exposure setting, the challenge is that it is unlikely that two units will have the same exact level of exposure, thus it is infeasible to create a finite sample representing a quasi-experimental arm with the same exposure level by solely matching on the generalized propensity score. Therefore, we propose a one-to-one nearest neighbor caliper matching procedure with replacement, which jointly matches both on the estimated generalized propensity score and exposure values. The idea behind our matching framework is that for each unit with exposure level w we find an observed unit that is both close to its exposure level, w, and its corresponding estimated generalized propensity score, $\hat{e}(w, \mathbf{c}_j)$ (see section 3.2 for details on how to estimate

e). The closeness of exposure level guarantees that the matched unit is a valid representation of observations for a particular exposure level, whereas, the closeness of generalized propensity score insures that we are properly adjusting for confounding.

The levels of closeness for both the exposure and generalized propensity score estimates need to be specified by distance measures. The generalized propensity score is a density function, thus there is no guarantee that the scale of the exposure and the generalized propensity score estimates are comparable. We standardize both quantities via a standardized Euclidean transformation, that is, $w_j^* = \frac{w_j - \min_j\ w_j}{\max_j\ w_j - \min_j\ w_j}$, $e^*(w_j, \mathbf{c}_j) = \frac{\hat{e}(w_j,\mathbf{c}_j) - \min_j\ \hat{e}(w_j,\mathbf{c}_j)}{\max_j\ \hat{e}(w_j,\mathbf{c}_j) - \min_j\ \hat{e}(w_j,\mathbf{c}_j)}$, where $\min_j$ and $\max_j$ are the minimum and maximum values across all observed units. Based on the standardized quantities, we propose a caliper metric matching function as

$$m_{\text{GPS}}(e, w) = \arg \min_{j: w_j \in [w-\delta, w+\delta]} ||(e^*(w_j, \mathbf{c}_j), w_j^*) - (e^*, w^*)||_{(\lambda, 1-\lambda)},$$

where $||.||_{(\lambda,1-\lambda)}$ is a pre-specified two-dimensional metric with weights $(\lambda, 1-\lambda)$. In practice, the metric can be defined as Manhattan Distance (ℓ_1 matching) or Euclidean Distance (ℓ_2 matching),

1. ℓ_1 matching: $m_{gps}(e, w) = \arg \min_{j: w_j \in [w-\delta, w+\delta]} \lambda \mid e^*(w_j, \mathbf{c}_j) - e^* \mid_1 + (1-\lambda) \mid w_j^* - w^* \mid_1$.

2. ℓ_2 matching: $m_{gps}(e, w) = \arg \min_{j: w_j \in [w-\delta, w+\delta]} \sqrt{\lambda \mid e^*(w_j, \mathbf{c}_j) - e^* \mid_2^2 + (1-\lambda) \mid w_j^* - w^* \mid_2^2}$.

The tuning parameter λ is introduced to control the relative weight that is attributed to the distance measures of the exposure versus the generalized propensity score estimates. The trade-off is between two source of bias, 1) the observed unit which was selected as representative of a target exposure level w does not have an exposure which is exactly w (rather is within a neighborhood), resulting in a bias estimate of mean potential outcome at w, 2) the observed unit does not match exactly on the target generalized propensity score value, resulting in a sacrifice of covariate balance in the matched dataset (Flores et al. 2007). Details on selecting λ are described in Section 3.4. The caliper δ is defined as the radius of the neighborhood around w, which means for any target exposure level w, we only allow for matches with an observed unit j satisfying $||W_j - w|| \leq \delta$. Details on how to select δ are discussed in Section 3.4.

3.2 Proposed Approach

Our proposed generalized propensity score matching approach contains two stages. The design stage in which we estimate the generalized propensity score, and the analysis stage in which we adjust for confounding by implementing the specified matching function based on the estimated generalized propensity score. The target estimands, the average causal exposure-response function, are then obtained.

1. Design Stage: For all units, estimate generalized propensity score via various parametric/non-parametric approaches. More specifically, fit a generalized propensity score model relating w to $\mathbf{C}$, $\hat{e}(w_j, \mathbf{c}_j) = \hat{g}^{-1}_{\mathbf{\Phi}}(\mathbf{c}_j)$, where g can be either parametric or non-parametric.

2. Analysis Stage: Define the suitable caliper matching function by specifying the desired metric, scale parameter λ, and caliper δ. Match individuals based on the matching function, that is a caliper metric matching proposed in Section 3.1. Impute $Y_j(w)$ as: $\hat{Y}_j(w) = Y^{obs}_{m_{\text{GPS}}(e(w,\mathbf{c}_j),w)}$ for $j = 1, \ldots, N$ successively. The matching estimator $\hat{\mu}(w)$ is equal to the overall average $\hat{E}[Y_j(w)]$ for each exposure level w.

3. Estimated average causal exposure-response function: Fit a normal-kernel smoother of $\hat{E}[Y_j(w)]$ on w to estimate the exposure-response function.

Hirano & Imbens (2004), Imai & Van Dyk (2004) relied on parametric models of the form $e(w, \mathbf{c}) = g^{-1}_{\mathbf{\Phi}}(\mathbf{c})$, where g is known, and compute the estimate $\hat{\mathbf{\Phi}}$ of $\mathbf{\Phi}$ by maximum likelihood estimation, such as a normal density,

$$\hat{e}(w_j, \mathbf{c}_j) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\big(- \frac{1}{2\pi\hat{\sigma}^2}(w_j - \hat{\eta}_0 - \hat{\boldsymbol{\eta}}_1^T \mathbf{c}_j)^2\big),$$

where $\hat{\eta}_0$, $\hat{\boldsymbol{\eta}}_1$, and $\hat{\sigma}$ are the maximum likelihood estimators. When parametric distributional assumptions are unlikely to hold, Flores et al. (2007), Galvao & Wang (2015) proposed flexible parametric/non-parametric models to estimate $g_{\mathbf{\Phi}}$, such as the Gaussian kernel estimator,

$$\hat{e}(w_j, \mathbf{c}_j) = n^{-1} h_r^{-(K+1)} \sum_{i=1}^{n} K\Big(\frac{w_j - w_i}{h_r}, \frac{\mathbf{c}_j - \mathbf{c}_i}{h_r}\Big)$$

where $K(.)$ is a Gaussian kernel density, and h_r is a selected bandwidth. Giné & Nickl (2008) showed that the kernel estimator converges weakly. Galvao & Wang (2015) argued that the non-parametric models have issues with its practical implementation, especially when there exist a large set of potential confounders, $\mathbf{C}_j$, which is likely in the context of air pollution studies, due to the curse of dimensionality. In general, one could choose any model specification relating w to $\mathbf{C}_j$ for its practice implementations.

3.3 Covariate Balance

In this section we introduce two measures of covariate balance; absolute correlation and blocked absolute standardized bias for continuous exposures. The absolute correlation between exposures and each pretreatment covariate is a global measure and can inform whether the whole matched dataset is balanced; while the blocked absolute standardized bias is estimated between $W_j \in [w-\delta, w+\delta]$ v.s. $W_j \notin [w-\delta, w+\delta]$ for every single exposure level w and is a local measure that informs which specific exposure levels are balanced. The block refers to the fact that the absolute standardized bias are calculated for W_j in the block $[w-\delta, w+\delta]$. The measures above build upon the work by Fong et al. (2018), Austin (2018) who examine covariate balance conditions with continuous exposures. We adapt them into the proposed matching framework.

Formally, we define $\{w_1 = w^0 + \delta, \ldots, w_I = w^0 + (2I-1)\delta\} \in [w^0, w^1]$, where $I = \lfloor \frac{w^1 - w^0}{2\delta} + \frac{1}{2} \rfloor$ is the number of blocks. Let m_i denote the number of units within the block $[w_i - \delta, w_i + \delta]$, where $i \in \{1, \ldots, I\}$. Suppose the k-th unit in the i-th block $[w_i - \delta, w_i + \delta]$ who has exposure W_{ik} and M-dimensional pretreatment covariates $\mathbf{C}_{ik}$ has outcome Y_{ik}, and it appears n_{ik} times in the matched dataset. We centralize and orthogonalize the covariates $\mathbf{C}_{ik}$ and the exposure W_{ik} as

$$\mathbf{C}^*_{ik} = \mathbf{S}_{\mathbf{C}}^{-1/2}(\mathbf{C}_{ik} - \bar{\mathbf{C}}_{ik}),\ W^*_{ik} = S_W^{-1/2}(W_{ik} - \bar{W}_{ik}),$$

where $\bar{\mathbf{C}}_{ik} = \sum_{i=1}^{I}\sum_{k=1}^{m_i} n_{ik}\mathbf{C}_{ik}/(N \cdot I)$, $\mathbf{S}_{\mathbf{C}} = \sum_{i=1}^{I}\sum_{k=1}^{m_i} n_{ik}(\mathbf{C}_{ik} - \bar{\mathbf{C}}_{ik})(\mathbf{C}_{ik} - \bar{\mathbf{C}}_{ik})^T/(N \cdot I)$, $W_{ik} = \sum_{i=1}^{I}\sum_{k=1}^{m_i} n_{ik}W_{ik}/(N \cdot I)$ and $S_W = \sum_{i=1}^{I}\sum_{k=1}^{m_i} n_{ik}(W_{ik} - \bar{W}_{ik})(W_{ik} - \bar{W}_{ik})^T/(N \cdot I)$.

Global Measure. Based on the global balancing condition, in a balanced population, we have the correlations between the exposure and pretreatment covariates are equal to zero, that is

$E[\mathbf{C}^*_{ik} W^*_{ik}] = \mathbf{0}$. We assess the covariate balance in the matched dataset as

$$\Big| \sum_{i=1}^{I} \sum_{k=1}^{m_i} n_{ik} \mathbf{C}^*_{ik} W^*_{ik} \Big| < \boldsymbol{\epsilon}_1,$$

in which each element of $\boldsymbol{\epsilon}_1$ is a pre-specified threshold, for example 0.1 (Zhu et al. 2015).

Local Measure. Based on the local balancing condition, the covariate balance for a specific exposure level can be defined by zero absolute standardized bias. We assess the covariate balance between units having exposure level within the block $[w_i - \delta, w_i + \delta]$ and outside of this block in the matched dataset as

$$\Big| \frac{\sum_{k=1}^{m_i} \mathbf{C}^*_{ik}}{N} - \frac{\sum_{i' \neq i} \sum_{k=1}^{m_{i'}} \mathbf{C}^*_{i'k}}{N \cdot (I-1)} \Big| < \boldsymbol{\epsilon}_2,$$

in which each element of $\boldsymbol{\epsilon}_2$ is a pre-specified threshold, for example 0.2 (Harder et al. 2010).

In addition, the average absolute correlations are defined as the average of absolute correlations among all covariates. The average blocked absolute standardized bias are defined as the average of absolute standardized bias among all covariates for each block.

1) average absolute correlation $\quad \Big|\Big| \sum_{i=1}^{I} \sum_{k=1}^{m_i} n_{ik} \mathbf{C}^*_{ik} W^*_{ik} \Big|\Big|_1 / M$

2) average absolute standardized bias $\quad \sum_{i=1}^{I} \Big|\Big| \frac{\sum_{k=1}^{m_i} \mathbf{C}^*_{ik}}{N} - \frac{\sum_{i' \neq i} \sum_{k=1}^{m_{i'}} \mathbf{C}^*_{i'k}}{N \cdot (I-1)} \Big|\Big|_1 / M$

3.4 Selecting the Tuning Parameters (λ, δ)

The optimal λ could be specified by minimizing a utility function that measures the degree of covariate balance measured by using with the global or local measures defined above. Noting that the optimal λ aim at achieving covariate balance on the global scale, the average absolute correlation would be the preferable measure in practice, moreover, it is more computational attainable. We implement data-driven procedures as follows: 1) construct the matched dataset using the matching function with a pre-specified $\lambda \in [0, 1]$, 2) calculate the desired covariate balance measure on the matched dataset, 3) Repeat steps 1-2 using grid search on λ, 4) select the λ that minimizes the

covariate balance measure. In practice, λ could be specified in the range $[0, 1]$ depending on the prioritization of adjusting for potential confounders and the potential effects heterogeneity due to different levels of exposures.

The caliper δ on the exposure is essential, since if we don't match the observed units within a neighborhood around the target exposure level w there is no guarantee that the matching estimator is unbiased. To achieve asymptotic properties for the matching estimator (Abadie & Imbens 2006), $\delta = O(N^{-1/2})$ is chosen which guarantees consistency and asymptotic normality. Theoretical justifications guide us to choose $\delta \approx \epsilon N^{-1/2}$, and in practice we set ϵ to be the range of the exposure levels, that is $\epsilon = \max_j w_j - \min_j w_j$. Unmatched units are those who do not have a good representative in the original dataset. If many units in $(w, \mathbf{c}_i)$, $(i = 1, \ldots, N)$, for a particular exposure level w are unmatched the causal exposure effect, $\mu(w)$, can not be estimated in the original dataset.

4 Asymptotic Properties

We present the asymptotic properties for the proposed caliper matching estimators for the average causal exposure-response function $\mu(w)$, where we match either on a scalar covariate, on the true generalized propensity score, or on the estimated generalized propensity score, given the caliper size $\delta = O(N^{-1/2})$ and $N\delta \to \infty$.

We begin by defining the conditional means and variances given covariates and given the generalized propensity score as follows:

$$\begin{aligned}
\mu_{\mathbf{C}}(w, \mathbf{c}) &= E\{Y_j(w) \mid W_j = w, \mathbf{C}_j = \mathbf{c}\} \\
\mu_{\text{GPS}}\{w, e(w, \mathbf{c})\} &= E\{Y_j(w) \mid W_j = w, e(w, \mathbf{C}_j) = e(w, \mathbf{c})\} \\
\sigma^2_{\mathbf{C}}(w, \mathbf{c}) &= Var\{Y_j(w) \mid W_j = w, \mathbf{C}_j = \mathbf{c}\} \\
\sigma^2_{\text{GPS}}\{w, e(w, \mathbf{c})\} &= Var\{Y_j(w) \mid W_j = w, e(w, \mathbf{C}_j) = e(w, \mathbf{c})\}
\end{aligned}$$

To simplify the problem, we only consider one-to-one nearest neighbor matching[1] on a set of

[1] According to the observation of Abadie & Imbens (2016), the number of matches per unit, M is small, often $M = 1$ in applications. Choosing a small M reduces finite sample biases caused by matching discrepancy through

continuous covariates $\mathbf{C}$. The matching estimator for $\mu(w)$ can be defined as,

$$\hat{\mu}(w) = \sum_{j=1}^{N} K(j) Y_j I_j(w, \delta)$$

where $K(j)$ indicates the number of replacements in which unit j is used as a match, and $I_j(w, \delta) = I_j([w-\delta, w+\delta])$. The difference between the matching estimator $\hat{\mu}(w)$, and the population average causal exposure-response function $\mu(w)$, can be decomposed as,

$$\hat{\mu}(w) - \mu(w) = \{\bar{\mu}(w) - \mu(w)\} + B_\mu(w) + \mathcal{E}_\mu(w) \tag{1}$$

where, $\bar{\mu}(w)$ is the average conditional means given covariates, $B_\mu(w)$ is the conditional bias of the matching estimator related to $\bar{\mu}(w)$, and $\mathcal{E}_\mu(w)$ is the average conditional residuals. Specifically, let $i(j)$ indicate the nearest neighbor match for unit $(w, \mathbf{C}_j)$,

$$\begin{aligned}
\bar{\mu}(w) &= \frac{1}{N}\sum_{j=1}^{N} \mu_{\mathbf{C}}(w, \mathbf{C}_j) \\
B_\mu(w) &= \frac{1}{N}\sum_{j=1}^{N} B_{\mu,j} = \frac{1}{N}\sum_{j=1}^{N} \{\mu_{\mathbf{C}}(W_{i(j)}, \mathbf{C}_{i(j)}) - \mu_{\mathbf{C}}(w, \mathbf{C}_j)\} \\
\mathcal{E}_\mu(w) &= \frac{1}{N}\sum_{j=1}^{N} K(j)\mathcal{E}_{\mu,j} I_j(w, \delta) = \frac{1}{N}\sum_{j=1}^{N} K(j)\{Y_j - \mu_{\mathbf{C}}(W_j, \mathbf{C}_j)\} I_j(w, \delta).
\end{aligned}$$

Lemma 3 (Matching Discrepancy) *Let* $j_1 = \underset{j=1,\ldots,N}{arg\,min} ||\mathbf{C}_j - \mathbf{c}||$ *and let* $U_1 = \mathbf{C}_{j_1} - \mathbf{c}$ *be the matching discrepancy. If* $\mathbf{C}$ *is scalar, then all the moments of* $N||U_1||$ *are uniformly bounded in* N.

Lemma 3 is the deduction of Lemma 2 presented in Abadie & Imbens (2006).

Theorem 1 (The Order of Bias) *Let* N_w *denote the number of units having exposures within the range of* $[w-\delta, w+\delta]$. *Assume Assumptions 1-3 and uniform boundedness Assumption (A1 in the Supplementary Material) hold, if* $\mathbf{C}$ *is scalar, the order of the bias of the proposed matching estimator, that is* $B_\mu(w)$, *is* $O_p\{(N\delta)^{-1}\}$.

larger values of M produce lower asymptotic variances. In the proposed continuous matching framework, we are prone to the matching discrepancy because of the restriction (caliper) on the units that can be matched, thus we tend to choose small M. All the asymptotic theories can be extended to one-to-M nearest neighbor matching.

Theorem 1 provides the stochastic order of bias terms in Equation 1. Under the described conditions, the bias term will be asymptotically negligible. Importantly, the rate is faster than $(N\delta)^{1/2}$, which guarantees the bias does not dominate the asymptotic behaviors of $\hat{\mu}(w)$.

Lemma 4 (Number of Replacement) *Assume Assumptions 1-3 hold, then* $K(j) = O_p(1/\delta)$, *and* $E[\{\delta K(j)\}^q]$ *is bounded uniformly in* N *for any* $q > 0$.

Lemma 4 is the extension of Theorem 3(i) presented in Abadie & Imbens (2006).

Theorem 2 (Variance) *Let* N_w *denote the number of units having exposures within the range of* $[w - \delta, w + \delta]$. *Assume Assumptions 1-3 and uniform boundedness assumption (A1 in the Supplementary Material) hold. If* $\mathbf{C}$ *is scalar,*

$$(N\delta)Var\{\hat{\mu}(w)\} = E\big[\sigma^2_{\mathbf{c}}(w, \mathbf{C}_j)\{\frac{3f_W(w)}{2e(w, \mathbf{C}_j)}\}\big] + o_p(1).$$

Theorem 2 shows the asymptotic variance for $\hat{\mu}(w)$ is finite, and provides an expression for it.

Theorem 3 (Consistency) *Assume Assumptions 1-3 and uniform boundedness assumption (A1 in the Supplementary Material) hold. If* $\mathbf{C}$ *is scalar,*

$$\hat{\mu}(w) - \mu(w) \to 0.$$

Theorem 3 is a key result, showing the proposed matching estimator is consistent.

Theorem 4 (Asymptotic Normality) *Assume Assumptions 1-3 and uniform boundedness assumption (A1 in the Supplementary Material) hold. If* $\mathbf{C}$ *is scalar,*

$$(N\delta)^{1/2}\{\hat{\mu}(w) - \mu(w)\} \to N\{0, \sigma_1^2(w)\}$$
$$\sigma_1^2(w) = E\big[\sigma^2_{\mathbf{c}}(w, \mathbf{C}_j)\{\frac{3f_W(w)}{2e(w, \mathbf{C}_j)}\}\big].$$

We show that when the set of matching covariates contains only one continuously distributed variable, the matching estimator is $(N\delta)^{1/2}$-consistent and asymptotic normal. Relative to matching

directly on the covariates, propensity score matching has the advantage of reducing the dimensionality of matching to a single dimension (Abadie & Imbens 2016). Therefore, for generalized propensity score matching, we have the following theorem.

Theorem 5 (Asymptotic Normality with Generalized Propensity Score) *Assume Assumptions 1-3 and uniform boundedness assumption (A2 in the Supplementary Material) hold.*

$$(N\delta)^{1/2}\{\hat{\mu}_{\text{GPS}}(w) - \mu(w)\} \to N\{0, \sigma_2^2(w)\}$$
$$\sigma_2^2(w) = E\big[\sigma_{\text{GPS}}^2\{w, e(w, \mathbf{C}_j)\}\{\frac{3f_W(w)}{2e(w, \mathbf{C}_j)}\}\big].$$

In practice, we rarely observe the true generalized propensity score values, and the generalized propensity score has to be estimated prior to matching. Abadie & Imbens (2016) proved and derived the large sample properties of propensity score matching estimators that corrects for the first step estimation of the propensity score. The main finding is that matching on the estimated propensity score has a smaller asymptotic variance than matching on the true propensity score when estimating average treatment effects.

Suppose our parametric model for the generalized propensity score is $e(w, \mathbf{c}) = g_{\mathbf{\Phi}}^{-1}(\mathbf{c})$, where g is known. We estimate $\hat{\mathbf{\Phi}}$ by maximum likelihood estimation. More specifically, we denote $\hat{\mu}_{\text{GPS}}(w; \hat{\mathbf{\Phi}})$ as the matching estimator with estimated generalized propensity score. We state the following theorem.

Theorem 6 (Asymptotic Normality with Estimated Generalized Propensity Score) *Assume Assumptions 1-3, uniform boundedness and almost sure convergence assumption (A2-3 in the Supplementary Material) hold.*

$$(N\delta)^{1/2}\{\hat{\mu}_{\text{GPS}}(w; \hat{\mathbf{\Phi}}) - \mu(w)\} \to N\{0, \sigma_2^2(w)\}$$

Theorem 6 states that no matter whether we match on the true generalized propensity score or the estimated generalized propensity score, the asymptotic properties are unchanged. Importantly, the asymptotic variance remains the same. The maximum likelihood estimation at the first step has convergence rate $N^{-1/2}$. As long as the first step estimation has convergence rate satisfying

$o_p\{1/(N\delta)^{-1/2}\}$, which also holds for many semi-/non-parametric estimations of the generalized propensity score, Theorem 6 holds. Proofs of Theorem 1-6 are provided in the Supplementary Material.

5 Simulations

5.1 Simulation Settings

We generate six confounders $(C_1, C_2, ..., C_6)$, which include a combination of continuous and categorical variables,

$$C_1, \ldots, C_4 \sim N(0, \boldsymbol{I}_4), C_5 \sim U\{-2, 2\}, C_6 \sim U(-3, 3),$$

and generate W using three specifications of the generalized propensity score model,

$$\begin{aligned}
&1) \;\; W = 9\{-0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)\boldsymbol{C}\} - 3 + N(0, 5) \\
&2) \;\; W = 15\{-0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)\boldsymbol{C}\} + 2 + 2T(4) \\
&3) \;\; W = 15\{-0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)\boldsymbol{C}\} + 3^{1/2}C_3^2 + T(4).
\end{aligned}$$

We generate Y from an outcome model which is assumed to be a cubical function of W with additive terms for the confounders and interactions between W and confounders,

$$\begin{aligned}
Y \mid W, \mathbf{C} &\sim N\{\mu(W, \mathbf{C}), 10\} \\
\mu(W, \mathbf{C}) &= -10 - (2, 2, 3, -1, 2, 2)\mathbf{C} - W(0.1 - 0.1C_1 + 0.1C_4 + 0.1C_5 + 0.1C_3^2) + 0.13^2W^3.
\end{aligned}$$

For these three specifications we vary the sample size $N(= 200, 1000, 5000)$ resulting in a total of nine scenarios. For each scenario, we generate 100 datasets.

After generating the data we estimate the exposure-response function for each simulation scenario using six different estimators including the proposed matching approach and three state-of-art alternatives, including two implementations of the Hirano and Imbens's generalized propensity score (HI-GPS) estimator (Hirano & Imbens 2004), inverse probability of treatment weighting

(IPTW) estimator (Robins et al. 2000), and two implementations of the doubly robust (DR) estimators (Bang & Robins 2005, Kennedy et al. 2017) (see the Supplementary Material for details). In addition, we estimate the exposure-response function based on the true data generating model as the gold standard. For implementation, some of these approaches rely on specification of the generalized propensity score model and/or outcome model. For those, unless otherwise specified we assume a linear generalized propensity score model, and consider a cubic function for the outcome model. To assess the performance of the different estimators, we calculate the absolute bias and mean squared error (MSE) of the estimated exposure-response function. These two quantities were estimated empirically at each point within range $\hat{\mathcal{W}}^*$, and integrated across the range $\hat{\mathcal{W}}^*$. Specifically, they are defined as follows:

$$\widehat{\text{Absolute Bias}} = \int_{\hat{\mathcal{W}}^*} |\frac{1}{S}\sum_{s=1}^{S} \hat{Y}_s(w) - Y(w)| f_W(w) dw$$

$$\widehat{\text{MSE}} = \int_{\hat{\mathcal{W}}^*} [\frac{1}{S}\sum_{s=1}^{S} \{\hat{Y}_s(w) - Y(w)\}^2]^{1/2} f_W(w) dw.$$

In the above $\hat{\mathcal{W}}^*$ denotes a trimmed version of the support of $\hat{\mathcal{W}}$, excluding 10% of mass at the boundaries.

5.2 Simulation Results

Simulation results are presented in Table 1. Not surprisingly, when the generalized propensity score model is correctly specified and does not have extreme values, the true outcome regression model (which is the cubical model) performs well, yet the HI-GPS estimator which relies on a linear regression model for the outcome is significantly biased due to the outcome model misspecification. On the other hand, when a correct outcome model structure (cubical) is assumed, HI-GPS (with a cubical regression model for the outcome), inverse probability of treatment weighting, and standard doubly robust estimators, all produce little bias and small MSE. Kennedy's doubly robust and the proposed matching estimator also perform well, and their absolute bias and MSEs decrease as the sample sizes increases. Yet in general, both approaches produce larger MSE compare to the other three estimators. Kennedy's doubly robust and our proposed matching estimators are both

non-parametric approaches which do not rely on outcome model specification and therefore are less efficient when the outcome model is correctly specified.

When the generalized propensity score model is correctly specified yet includes extreme generalized propensity score values, the inverse probability of treatment weighting estimator and standard doubly robust estimator produce extremely large MSE, and are not able to reduce confounding bias. This is not surprising as they rely on weighting and are sensitive to extreme generalized propensity score values (Waernbaum 2012). However, Kennedy's non-parametric doubly robust estimator has better performance. The HI-GPS estimator only performs well in this setting when the outcome model is correctly specified as a cubical model. Our proposed matching estimator outperforms HI-GPS, inverse probability of treatment weighting and the standard doubly robust estimator, both in terms of absolute bias and MSE, and achieves performance as good as Kennedy's state-of-the-art doubly robust method.

When the generalized propensity score model is misspecified, the proposed matching approach provides bias reduction and smaller MSE compared to all other approaches. The result is not surprising, since 1) HI-GPS and the inverse probability of treatment weighting estimator require correctly specifying the generalized propensity score model, which is not the case here, and are sensitive to model misspecifications, 2) as discussed in Waernbaum (2012), although standard doubly robust robust approaches guarantee appealing large-sample properties when at least one model (either the generalized propensity score or outcome model) is correctly specified, they can perform very poorly in finite-sample cases if the generalized propensity score model is misspecified, which is the case in this simulation setting, 3) Kennedy's doubly robust estimator performs well since it does not require parametric assumptions, and overcomes the model misspecification by flexible machine learning, yet in the settings of existing extreme weights, it is less precise than matching.

In general, via simulations, we see that the proposed matching approach outperforms existing methods under settings of model misspecifications and/or in presence of extreme generalized propensity score values. This improved performance is due to the fact that our proposed matching approach does not require any parametric assumptions for the outcome model and is more robust to misspecification of the generalized propensity score model compared to covariate adjustment

and weighting-type approaches.

5.3 Covariate Balance Assessment

The matching framework provides a transparent way to assess covariate balance. In practice, we can compare values of covariate balance measures, for example absolute correlations, described in Section 3.3, between the matched dataset and original dataset. If the absolute correlations for each of the covariates in the matched dataset are substantially smaller than those in the original dataset, we conclude that our approach improves covariate balance. Moreover, Zhu et al. (2015) suggests that confounding between the exposure and the outcome is small when the average absolute correlations are less than 0.1.

Figure 1 presents absolute correlation results from three simulation settings where we vary the specification of the generalized propensity score model, under sample size $N = 5000$. We assess balance by calculating the absolute correlation for each of six covariates. We see that balance improves substantially across all six covariates for all three simulation settings. Under the setting where the generalized propensity score model is correctly specified without extreme generalized propensity score values (scenario 1), absolute correlations for all confounders are ≤ 0.10, which indicates little imbalance within the matched dataset. Under the setting of extreme values for the generalized propensity score (scenario 2) and/or misspecification of generalized propensity score model (scenario 3), the proposed matching procedure improves balance for all covariates, though there is still evidence of imbalance.

6 Data Application

We apply the proposed matching method to estimate the effect of long-term $PM_{2.5}$ exposure on all-cause mortality, using information on Medicare participants across New England (VT, NH, CT, MA, RI and ME) from 2000 to 2012. This study population includes a total of 3.3 million individuals with 24.5 million person-years of follow up, who reside in 2,202 zip codes. We constructed counts corresponding to the number of deaths and mortality rates were based on the total number of person-years for Medicare enrollees for each zip code per year across New England. $PM_{2.5}$

exposures were determined at 1km $\times$ 1km grid cells using a spatio-temporal prediction model (Di et al. 2016). Medicare data are available at the zip code level, yet $PM_{2.5}$ exposures are estimated at the grid level. To obtain annual average $PM_{2.5}$ at each zip code, we aggregate these gridded concentrations by taking area-weighted averages.

Design Stage. We assume that the generalized propensity score model is a linear regression model with the exposure specified as the dependent variable and 20 potential confounders including population demographic information, weather information, individual-level information, and temporal trends (calendar year), as predictors.

Analysis Stage. We implement our proposed matching approach using the estimated generalized propensity score. Specifically, we use a pre-specified ℓ_1 matching function with scale parameter $\lambda = 0.2$ and caliper $\delta = 0.05$. The choice of (λ, δ) follows the guidance in Section 3.4. After matching, we fit the normal-kernel smoother to estimate the causal exposure-response function relating average potential outcome within each caliper to the corresponding $PM_{2.5}$ levels.

We assess covariate balance by calculating the absolute correlation for each potential confounder as discussed in Section 3.3. The generalized propensity score implementation largely improves covariate balance for 15 out of 20 potential confounders. The average absolute correlation is 0.25 before matching, whereas, the average absolute correlation is 0.07 after matching (See Figure 2). Importantly, time trend (calender year) that had a strong imbalance before matching is balanced after matching.

Figure 3 shows the average causal exposure-response function. We found an approximate linear causal relationship between mortality and long-term $PM_{2.5}$ exposures across the whole range of annual average $PM_{2.5}$ ($2.05 - 15.43$ $\mu g/m^3$) in New England. Although the current long-term $PM_{2.5}$ exposure standard is an annual mean of 12.0 $\mu g/m^3$, refer to National Ambient Air Quality Standards (NAAQS) Table (USEPA 2012), there is an increasing interest in studying the effect of $PM_{2.5}$ exposures at lower levels (Villeneuve et al. 2015, Shi et al. 2016, Di et al. 2017). Our results are consistent with recent epidemiological studies which have found a strong association between long-term exposure to $PM_{2.5}$ and mortality at low exposure levels. Our approach provides a causal interpretation of the effects of long-term $PM_{2.5}$ exposure on all-cause mortality at lower exposure levels; using our proposed matching approach we found that each 1 $\mu g/m^3$ increase in

annual average $PM_{2.5}$ exposure causes an approximately 7.4×10^{-4} increase in all-cause mortality hazard. This causal effect estimate is consistent with the findings in Wang et al. (2017).

7 Discussion

We develop an innovative approach for estimating causal effects using observational data in settings with continuous exposures, and introduce a new framework for generalized propensity score caliper matching. Our proposed approach fills an important gap in the literature as it provides a theoretically justified generalization from Abadie & Imbens (2006, 2016) for matching in the context of continuous exposures. We also demonstrated that under the local weak unconfoundedness assumption, the newly proposed matching estimators attain $(N\delta)^{1/2}$-consistency and asymptotic normality if the caliper δ is well chosen. By conducting simulation studies with a wide range of data generating mechanisms, we found that the proposed matching framework shares advantages that have been previously discussed in literature (Rosenbaum & Rubin 1983, Ho et al. 2007, Zubizarreta 2012, Waernbaum 2012), including that 1) it is robust to misspecification of the generalized propensity score model, especially in the presence of extreme values (Waernbaum 2012), 2) it does not depend on the specification of the outcome model, 3) it is straightforward to assess covariate balance. In addition, we adapt two covariate balance measures into the proposed matching framework, and describe the way to assess balance based on these measures.

The proposed approach relies on four main assumptions 1) consistency, 2) overlap, 3) local weak unconfoundedness, and 4) smoothness. The consistency assumption is a fundamental assumption in the classical potential outcome framework. Recent literature (Tchetgen & VanderWeele 2012, Papadogeorgou et al. 2017) relaxes this assumption by allowing interference, yet future work is warranted to combine this concept with (generalized) propensity score-based analyses. The overlap assumption is another fundamental assumption for the validity of most causal inference methods. Under binary or categorical exposure cases, investigators widely use diagnostic plots to check overlap (Braun et al. 2017, Wu et al. 2017) and trimming techniques to ensure overlap (Crump et al. 2009, Harder et al. 2010, Yang et al. 2016). However, when the exposure of interest is a continuous variable, the overlap is defined by a density function on a Lebesgue set, and therefore it is conceptually difficult to check it directly via finite samples. One could categorize the continuous exposure

and check/ensure overlap among categories using standard approaches developed in categorical exposure cases (Yang et al. 2016, Wu et al. 2017), yet no current approach allows verification of overlap on the continuous scale. In future work it will be useful to study rigorous approaches for checking and ensuring overlap in this setting.

We introduced the local weak unconfoundedness assumption which is less stringent than the common weak unconfoundedness assumption, though it is still unverifiable since data are always uninformative about the distribution of the counterfactual outcome. As with other (generalized) propensity score-based analyses, the matching approach does not resolve potential bias due to unmeasured confounding, in which case the unconfoundedness assumption is violated. By choosing a suitable degree of local approximation, that is selecting δ in the local weak unconfoundedness assumption, we can find in theory when the proposed matching estimator achieves desirable asymptotic properties. The smoothness assumption is essentially the standard smoothness condition imposed in non-parametric regression problems. In addition, we require the rate of smoothness, the bandwidth, to satisfy $h \geq \delta$ as it guarantees that the outcome does not jump dramatically within the neighborhood of an exposure level. It is feasible to choose (δ, h) that achieve desirable asymptotic properties, yet it would be helpful to develop new methods to search for optimal (δ, h) via observed data as part of future work. Although in this paper, we focus on one-to-one nearest neighbor matching, all the asymptotic theories can be extended to one-to-M nearest neighbor matching. However, as discussed in Abadie & Imbens (2016), in applications, M is usually small, often $M = 1$, since choosing a small M reduces finite sample biases, in contrast to larger values of M which produce lower asymptotic variances. One important advantage of our proposed matching framework is that it aims at constructing a matched dataset that mimics a randomized experiment, thus making it feasible to identify/estimate causal quantities other than the average causal exposure-response function, such as quantile causal exposure-response function.

We applied the approach to estimate the causal effect of long-term $PM_{2.5}$ exposure on all-cause mortality. Previous air pollution health studies that used propensity score-based analyses dichotomized or categorized continuous exposure variables to utilize propensity score methods (Baccini et al. 2017, Wu et al. 2017). The generalized propensity score caliper matching approach introduced in this paper is the first matching approach which allows for estimating causal effects on

continuous exposures and assesses covariate balance in a straightforward way. The simplicity and generalizability of our matching framework can promote awareness of and interest in estimating causality in future applied research, especially in fields where interventions/exposures/treatments are naturally continuous such as economics, political science, and environmental health.

Acknowledgement

The authors are grateful to José R. Zubizarreta and Elizabeth A. Stuart for helpful discussions. Funding was provided by the Health Effects Institute (HEI) grant 4953-RFA14-3/16-4, National Institute of Health (NIH) grants R01 GM111339, R01 ES024332, R01 ES026217, R01 ES028033, R01 MD012769, DP2 MD012722, NIH/National Institute on Minority Health and Health Disparities (NIMHD) grant P50 MD010428, and USEPA grants RD-83587201-0, RD-83615601. The contents are solely the responsibility of the grantee and do not necessarily represent the official views of the funding agencies. Further, funding agencies do not endorse the purchase of any commercial products or services related to this publication. Research described in this article was conducted under contract to the HEI, an organization jointly funded by the USEPA (Assistance Award No. R-83467701) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers. The computations in this paper were run on the Research Computing Environment (RCE) supported by the Institute for Quantitative Social Science in the Faculty of Arts and Sciences at Harvard University.

References

Abadie, A. & Imbens, G. W. (2006), 'Large sample properties of matching estimators for average treatment effects', *econometrica* **74**(1), 235–267.

Abadie, A. & Imbens, G. W. (2016), 'Matching on the estimated propensity score', *Econometrica* **84**(2), 781–807.

Austin, P. C. (2018), 'Assessing covariate balance when using the generalized propensity

score with quantitative or continuous exposures', *Statistical Methods in Medical Research* p. 0962280218756159.

Baccini, M., Mattei, A., Mealli, F., Bertazzi, P. A. & Carugno, M. (2017), 'Assessing the short term impact of air pollution on mortality: a matching approach', *Environmental Health* **16**(1), 7.

Bang, H. & Robins, J. M. (2005), 'Doubly robust estimation in missing data and causal inference models', *Biometrics* **61**(4), 962–973.

Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z. J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M. et al. (2014), 'Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 european cohorts within the multicentre escape project', *The Lancet* **383**(9919), 785–795.

Braun, D., Gorfine, M., Parmigiani, G., Arvold, N. D., Dominici, F. & Zigler, C. (2017), 'Propensity scores with misclassified treatment assignment: a likelihood-based adjustment', *Biostatistics* p. kxx014.

Cao, W., Tsiatis, A. A. & Davidian, M. (2009), 'Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data', *Biometrika* **96**(3), 723–734.

Crump, R. K., Hotz, V. J., Imbens, G. W. & Mitnik, O. A. (2009), 'Dealing with limited overlap in estimation of average treatment effects', *Biometrika* **96**(1), 187–199.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y. & Schwartz, J. (2016), 'Assessing pm2. 5 exposures with high spatiotemporal resolution across the continental united states', *Environmental science & technology* **50**(9), 4712–4721.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F. & Schwartz, J. D. (2017), 'Air pollution and mortality in the medicare population', *New England Journal of Medicine* **376**(26), 2513–2522.

Flores, C. A. et al. (2007), 'Estimation of dose-response functions and optimal doses with a continuous treatment', *University of Miami. Typescript* .

Fong, C., Hazlett, C., Imai, K. et al. (2018), 'Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements', *The Annals of Applied Statistics* **12**(1), 156–177.

Galvao, A. F. & Wang, L. (2015), 'Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment', *Journal of the American Statistical Association* **110**(512), 1528–1542.

Giné, E. & Nickl, R. (2008), 'Uniform central limit theorems for kernel density estimators', *Probability Theory and Related Fields* **141**(3-4), 333–387.

Harder, V. S., Stuart, E. A. & Anthony, J. C. (2010), 'Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research.', *Psychological methods* **15**(3), 234.

Hirano, K. & Imbens, G. W. (2004), 'The propensity score with continuous treatments', *Applied Bayesian modeling and causal inference from incomplete-data perspectives* **226164**, 73–84.

Ho, D. E., Imai, K., King, G. & Stuart, E. A. (2007), 'Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference', *Political analysis* **15**(3), 199–236.

Imai, K. & Van Dyk, D. A. (2004), 'Causal inference with general treatment regimes: Generalizing the propensity score', *Journal of the American Statistical Association* **99**(467), 854–866.

Imbens, G. W. (2000), 'The role of the propensity score in estimating dose-response functions.', *Biometrika* **87**(3).

Imbens, G. W. & Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Kang, J. D., Schafer, J. L. et al. (2007), 'Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data', *Statistical science* **22**(4), 523–539.

Kennedy, E. H., Ma, Z., McHugh, M. D. & Small, D. S. (2017), 'Non-parametric methods for doubly robust estimation of continuous treatment effects', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(4), 1229–1245.

Kioumourtzoglou, M.-A., Schwartz, J., James, P., Dominici, F. & Zanobetti, A. (2016), 'Pm2. 5 and mortality in 207 us cities: modification by temperature and city characteristics', *Epidemiology* **27**(2), 221–227.

Papadogeorgou, G., Mealli, F. & Zigler, C. (2017), 'Causal inference for interfering units with cluster and population level treatment allocation programs', *arXiv preprint arXiv:1711.01280* .

Rassen, J. A., Shelat, A. A., Franklin, J. M., Glynn, R. J., Solomon, D. H. & Schneeweiss, S. (2013), 'Matching by propensity score in cohort studies with three treatment groups', *Epidemiology* **24**(3), 401–409.

Robins, J. M., Hernan, M. A. & Brumback, B. (2000), 'Marginal structural models and causal inference in epidemiology'.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American statistical Association* **89**(427), 846–866.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* pp. 41–55.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of educational Psychology* **66**(5), 688.

Rubin, D. B. et al. (2008), 'For objective causal inference, design trumps analysis', *The Annals of Applied Statistics* **2**(3), 808–840.

Shi, L., Zanobetti, A., Kloog, I., Coull, B. A., Koutrakis, P., Melly, S. J. & Schwartz, J. D. (2016), 'Low-concentration pm2. 5 and mortality: Estimating acute and chronic effects in a population-based study', *Environmental health perspectives* **124**(1), 46.

Tchetgen, E. J. T. & VanderWeele, T. J. (2012), 'On causal inference in the presence of interference', *Statistical methods in medical research* **21**(1), 55–75.

USEPA (2012). US Environmental Protection Agency. National Ambient Air Quality Standards (NAAQS) Table: https://www.epa.gov/criteria-air-pollutants/naaqs-table.

Villeneuve, P. J., Weichenthal, S. A., Crouse, D., Miller, A. B., To, T., Martin, R. V., van Donkelaar, A., Wall, C. & Burnett, R. T. (2015), 'Long-term exposure to fine particulate matter air pollution and mortality among canadian women', *Epidemiology* **26**(4), 536–545.

Waernbaum, I. (2012), 'Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation', *Statistics in medicine* **31**(15), 1572–1581.

Wang, Y., Kloog, I., Coull, B. A., Kosheleva, A., Zanobetti, A. & Schwartz, J. D. (2016), 'Estimating causal effects of long-term $pm_{2.5}$ exposure on mortality in New Jersey', *Environmental health perspectives* **124**(8), 1182.

Wang, Y., Lee, M., Liu, P., Shi, L., Yu, Z., Awad, Y. A., Zanobetti, A. & Schwartz, J. D. (2017), 'Doubly robust additive hazards models to estimate effects of a continuous exposure on survival', *Epidemiology (Cambridge, Mass.)* **28**(6), 771.

Wu, X., Braun, D., Kioumourtzoglou, M.-A., Choirat, C., Di, Q. & Dominici, F. (2017), 'Causal inference in the context of an error prone exposure: air pollution and mortality', *arXiv preprint arXiv:1712.00642* .

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E. & Kadziola, Z. (2016), 'Propensity score matching and subclassification in observational studies with multi-level treatments', *Biometrics* **72**(4), 1055–1065.

Zhu, Y., Coffman, D. L. & Ghosh, D. (2015), 'A boosting algorithm for estimating generalized propensity scores with continuous treatments', *Journal of causal inference* **3**(1), 25–40.

Zubizarreta, J. R. (2012), 'Using mixed integer programming for matching in an observational study of kidney failure after surgery', *Journal of the American Statistical Association* **107**(500), 1360–1371.

Table 1: Absolute Bias and Mean Squared Error (MSE) when the true outcome model is a cubical function of the exposure, based on 100 simulation replicates

GPS model fit	N	**Matching**	HI-GPS Cubical	HI-GPS Linear	IPTW	Standard DR	Kennedy's DR	True Reg
1) Correctly	200	2.05 (3.72)	0.73 (2.02)	9.56 (10.86)	0.27 (1.98)	0.09 (1.70)	0.70 (2.87)	0.07 (1.19)
Specified, No	1000	1.06 (2.17)	0.76 (1.18)	9.52 (9.84)	0.16 (1.04)	0.28 (3.78)	0.57 (2.17)	0.05 (0.53)
Extreme GPS	5000	0.51 (1.18)	0.80 (0.94)	9.47 (9.54)	0.08 (0.59)	0.09 (0.54)	0.34 (1.41)	0.01 (0.24)
2) Correctly	200	2.13 (4.69)	1.70 (3.44)	15.32 (21.05)	0.57 (7.64)	$*(*)$	1.50 (6.01)	0.16 (1.94)
Specified,	1000	1.69 (3.58)	1.63 (2.09)	15.92 (17.73)	80.50 $(*)$	$*(*)$	0.93 (3.39)	0.05 (0.80)
Extreme GPS	5000	0.89 (2.17)	1.67 (1.78)	16.32 (19.12)	$*(*)$	$*(*)$	1.13 (3.39)	0.02 (0.33)
	200	1.86 (4.66)	1.84 (3.92)	14.56 (16.45)	3.45 (10.38)	$*(*)$	1.72 (4.53)	0.13 (3.25)
3) Misspecified	1000	0.89 (3.24)	1.85 (2.47)	14.81 (15.31)	12.54 (74.31)	$*(*)$	0.81 (2.61)	0.07 (1.38)
	5000	0.69 (2.26)	2.07 (2.21)	14.86 (15.03)	33.86 $(*)$	$*(*)$	0.86 (2.71)	0.03 (0.56)

Notes: Matching = the proposed generalized propensity score caliper matching; HI-GPS Cubical = HI-GPS that adds generalized propensity score as covariates in a cubical outcome model; HI-GPS Linear = HI-GPS that adds generalized propensity score as covariates in a linear outcome model; IPTW = inverse probability of treatment weighting; Standard DR = doubly robust estimator that uses parametric models (Bang & Robins 2005); Kennedy's DR = doubly robust estimator that uses non-parametric models (Kennedy et al. 2017); True Reg = regression using true model. Additional details on the various approaches can be found in the Supplementary Material. * represents values larger than 100.

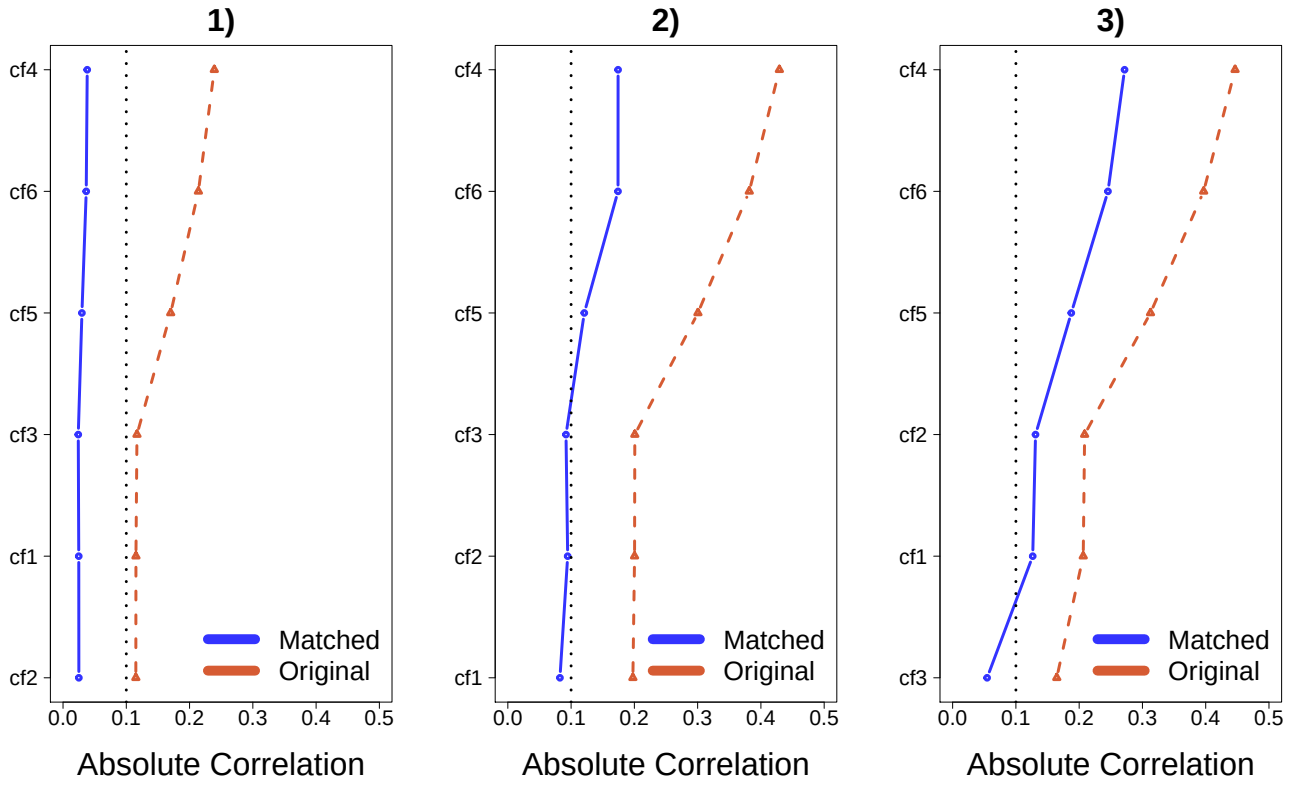

Figure 1: Absolute Correlations (ACs). Each panel represents the ACs for each covariate in the matched dataset (solid line) and original dataset (dashed line) under three simulation settings where generalized propensity score model specifications vary. The dotted line represents the cut-off of covariate balance suggested by Zhu et al. (2015). generalized propensity score matching improves covariate balance for all six covariates in all settings.

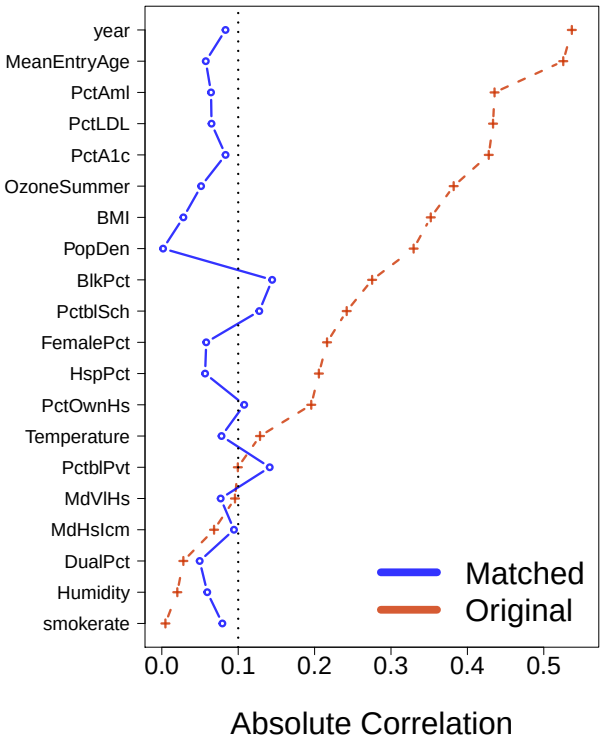

Figure 2: Absolute Correlations (ACs). The figure represents the ACs for each covariate in the matched dataset (solid line) and original dataset (dashed line). The dotted line represents the cut-off of covariate balance suggested by Zhu et al. (2015). In general, generalized propensity score matching substantially improves covariate balance for these potential confounders. The average AC is 0.25 before matching, and 0.07 after matching. Importantly, time trend (year) has a strong imbalance before matching, yet is balanced after matching.

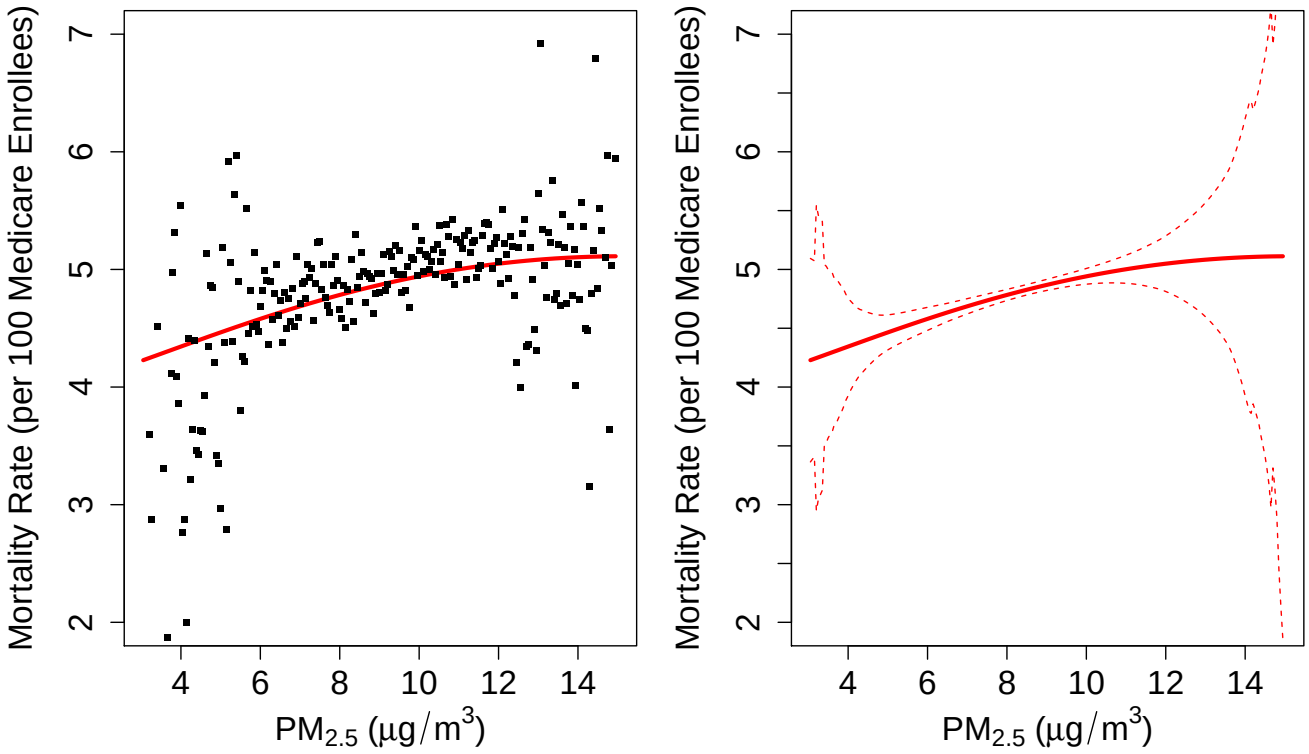

Figure 3: The causal exposure-response function relating all-cause mortality to long-term $PM_{2.5}$ exposure. The left panel presents the estimated average potential outcome within each caliper (black dots) along with the curve obtained by fitting a normal-kernel smoother with optimal bandwidth (red solid line). The right panel is the smoothed curve with its point-wise confidence band calculated by m-out-of-n bootstrap.