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 19, 2020

Abstract

The robustness of statistical methods and interpretability of statistical analysis from observational studies are central concerns in many applied fields, and robust causal inference methods have the potential to mitigate these concerns. This is particularly crucial in air pollution studies where the findings inform key political and policy decisions. When estimating the causal effects of continuous exposure (i.e., air pollution levels) in observational studies, generalized propensity scores (GPS) have been used to adjust for confounding bias. Existing GPS methods, either relying on weighting or regression, have certain limitations: a) they require a correctly specified outcome model; b) they are sensitive to extreme values of the estimated GPS; c) assessing covariate balance when using these approaches is not straightforward. Matching, a class of popular causal inference methods with binary treatments, has not been extended to the continuous exposure setting, disregarding its many attractive features on method robustness and interpretability. In this paper, we propose an innovative approach for GPS caliper matching 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 mild smoothness conditions, we provide theoretical guarantees that our proposed matching estimators attain consistency and asymptotic normality. Importantly, we introduce new measures of covariate balance under the matching framework. In simulations, our proposed matching estimator outperforms existing methods under settings of model misspecification and/or in the presence of extreme values of the estimated GPS in terms of bias reduction, root mean squared error, and overall achieves excellent covariate balance. We utilize the largest-to-date Medicare claims data for the entire US from 2000 to 2016 to construct a continuous causal exposure-response curve for long-term exposure to fine particles ($PM_{2.5}$) on mortality. We observed aggravated harmful effects at exposure levels below the national standards, and suggest that a revision to lower the current standards is indispensable for improving public health.

Keywords: Causal Inference, Continuous Treatment, Covariate Balance, Non-parametric, Observational Study

1 Introduction

Estimating the causal effects in many air pollution epidemiology studies is challenging as 1) the treatment (or named exposure in this setting) is continuous in nature, thus one has to allow for flexible estimation of the exposure-response function on a continuous scale; 2) they are observational studies, thus often contain a large set of covariates that are associated with both the exposure and outcome of interest (potential confounders); and 3) these studies are often large-scale and are therefore computationally burdensome requiring the analysis of massive datasets containing millions of observations.

In most air pollution epidemiology studies confounding adjustment is achieved by fitting a regression model relating the outcome to the exposures and covariates (e.g., Di et al. (2017), Liu et al. (2019), Wang et al. (2019)). These conventional regression methods mix the design and analysis stages, are susceptible to model misspecification, offer limited sensitivity analysis tools, and therefore allow for likely deviations from causality (Rubin et al. 2008, Imai et al. 2018). Among ongoing debates on air pollution policy (Dominici et al. 2014, Goldman & Dominici 2019, Peters et al. 2019), there is a need for more robust and interpretable causal inference methodology (Carone et al. 2019, Bind 2019). Under a potential outcomes framework for causal inference, the design stage (where we aim at defining suitable estimands and achieving covariate balance to adjust for confounding bias) and the analysis stage (where we aim at estimating causal effects) are distinct (Imbens & Rubin 2015). A common approach for confounding adjustment in this framework is using propensity scores, i.e., the probability of a unit being assigned to a particular exposure, given the pre-exposure covariates. Using propensity scores to adjust for confounding in a causal inference framework was first introduced by Rosenbaum & Rubin (1983). After this seminal paper, advanced propensity score techniques, both for estimation and implementation, have been developed to estimate causal effects in observational studies, as reviewed by 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 for a categorical exposure. To handle categorical exposures, Imbens (2000) developed the generalized propensity scores (GPS), a natural analogue to propensity score estimation which uses multinomial logistic regression instead of logistic regression to predict multiple exposure levels for the propensity score model. Imbens (2000)

describe an analogue for implementing categorical GPS using inverse probability of treatment weighting (IPTW). Although there is no natural analogue for matching and subclassification for categorical GPS (Rassen et al. 2013), Yang et al. (2016) propose an alternative way to estimate causal effects using matching and subclassification in this setting.

Hirano & Imbens (2004) propose an extension of the categorical GPS to continuous exposures. Yet, Hirano & Imbens (2004) conduct both the estimation and inference assuming correctly parametric specifications of both the GPS and outcome models by including the estimated GPS values as a covariate in the outcome model. Robins et al. (2000) propose an approach using marginal structure models in which the parameters can be consistently estimated using a class of IPTW estimators. However, this approach also requires the correct specification of both GPS and outcome models. A doubly robust estimator first proposed by Robins et al. (1994), Bang & Robins (2005), is a class of augmented IPTW 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 models of the (generalized) propensity score and the outcome are misspecified, which is likely to happen in real-world 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 gaining model robustness in both GPS and outcome models.

The matching estimation, another class of popular causal inference approaches in binary and categorical settings (Rosenbaum & Rubin 1983, Yang et al. 2016), has the following attractive features: 1) it is more robust to misspecifications of the GPS model, especially in the presence of extreme values of GPS (Waernbaum 2012); 2) it is completely free of parametric model assumptions of outcome analyses (Lu et al. 2011, Zubizarreta 2012); and 3) it maintains the unit of analysis intact and thus allows for the transparent assessment of covariate balance, secondary and sensitivity analyses (Zubizarreta 2012, Rosenbaum 2020). Yet, the matching estimator has never been extended and implemented in causal inference for continuous exposures. Indeed, at present no natural extension of matching exists that accommodates the continuous feature of the exposures.

In our work, we focus on settings for which we have a continuous exposure and propose a novel

GPS caliper matching framework that jointly matches on both the estimated GPS and exposure levels to fully adjust for confounding bias. Our motivating application is to analyze the causal effects of long-term $PM_{2.5}$ exposure on mortality in a massive observational cohort constructed by all Medicare beneficiaries ($n = 68,503,979$) in the continental United States (2000–2016). Medicare claims data, obtained from the Centers for Medicare and Medicaid Services (CMS), provide a rich data platform to conduct air pollution studies on a national scale (Di et al. 2017). Our proposed approach aims at gaining method robustness and interpretability in both the design and analysis stages to credibly assess causal effects in observational studies.

2 The Generalized Propensity Score Function

We used the following mathematical notation: let N denote the study sample size. For each unit $j \in \{1, \ldots, N\}$, let $\mathbf{C}_j$ denote the pre-exposure covariates for unit j, which is characterized by a q-vector $(C_{1j}, \ldots, C_{qj})$; 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 $p_j\{W \mid C, Y(w)\}$, for all $w \in \mathbb{W}$, denote the assignment mechanism defined as the conditional probability 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 pre-exposure covariates $\mathbf{c}$, we will be able to estimate $\mu(w)$ 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 defined above refers to the fact that we do not require (conditional) independence of potential outcomes, $Y(w)$, 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 now introduce Assumption 3, the Local Weak Unconfoundedness assumption, which is less stringent than the weak unconfoundedness assumption defined above.

Assumption 3 (Local Weak Unconfoundedness) *Let $I_j(.)$ be the 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 it follows a positive sequence tending to zero as $N \to \infty$. The assignment mechanism is local weakly unconfounded if for all $w \in \mathbb{W}$, in which w is continuously distributed with respect to the Lebesgue measure on $\mathbb{W} = [w^0, w^1]$, 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 a 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 $h \simeq \delta$, 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 pre-exposure 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) the local weak unconfoundedness holds when we condition on the GPS, 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 GPS) *Suppose the assignment mechanism is local 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 local weakly unconfounded. Then for all* $w \in \mathbb{W} = [w^0, w^1]$,

$$\mu(w) = E[Y_j(w)] = \lim_{\delta \to 0} E\big[E(Y_j^{obs} | e(w_j, \mathbf{C_j}), w_j \in [w-\delta, w+\delta]]\big].$$

Even though the causal exposure effects are constructed by conditioning on a single realization of the GPS instead of by conditioning on the full set of GPS, under the local weak unconfoundedness assumption, the estimated population average causal exposure-response function is still identifiable (Hirano & Imbens 2004). The proofs of both Lemmas are presented in the Supplementary Materials.

3 Matching Framework

3.1 General matching function

The ultimate objective for matching is to construct matched datasets that approximate a randomized experiment as closely as possible by achieving good covariate balance. In the categorical setting, Yang et al. (2016) proposed a GPS 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 exact same 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 GPS. Therefore, we propose a one-to-M[1] nearest neighbor caliper matching procedure with replacement, which jointly matches

[1]According to 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 discrepancies, though larger values of M produce lower asymptotic variances.

both on estimated GPS 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 GPS, $\hat{e}(w, \mathbf{c_j})$ (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 GPS insures that we are properly adjusting for confounding.

The levels of closeness for both the exposure and GPS estimates need to be specified by distance measures. The GPS is a density function, thus there is no guarantee that the scale of exposure and the GPS 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 the specification of the matching function, that is a caliper metric matching as

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

where $||\ .\ ||$ is a pre-specified two-dimensional metric, for example, Manhattan Distance (L1 matching) or Euclidean Distance (L2 matching). (λ, δ) are named the scale parameter and the caliper respectively. Details on the role and selection of (λ, δ) are discussed in Section 3.4.

3.2 Proposed Approach

Our proposed GPS matching approach contains two key stages: a) design and b) analysis. Aligning with Stuart (2010), the design stage deliberately prevents the access to the outcome information in the study, designing the observational study as if it were a randomized experiment. Only after the design, does the analysis stage begin, involving the outcomes to estimate the average causal exposure-response function.

a) Design Stage:

1) For all units, estimate GPS via various parametric/non-parametric approaches. More specifically, fit a GPS model relating W to $\mathbf{C}$, $\hat{e}(w_j, \mathbf{c}_j) = g^{-1}_{\hat{\mathbf{\Phi}}}(\mathbf{c}_j)$. Various flexible parametric/non-parametric specifications of g were proposed in Hirano & Imbens (2004), Imai & Van Dyk (2004), Flores et al. (2007), Van der Laan et al. (2007), Galvao & Wang (2015), Kennedy et al. (2017).

2) Define the suitable caliper matching function proposed in Section 3.1 by specifying the desired metric, scale parameter λ, and caliper δ. Specifically, both (λ, δ) can be seen as tuning parameters and selected based on data-driven procedures proposed in Section 3.4.

3) Match individuals based on the specified caliper metric matching function. 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. We constructed the matched data by collected imputed $\hat{Y}_j(w)$ for all predetermined levels of $w \in [w^0, w^1]$.

4) Assess covariate balance for each pre-exposure covariates on matched data by using measures proposed in Section 3.3.

b) Analysis Stage:

5) After implementing a suitable matching in the deign stage and constructing the matched data, the simple matching estimator $\hat{\mu}(w) = \hat{E}\{\hat{Y}_j(w)\}$ can be obtained for all predetermined exposure levels w.

6) Obtaining a smoothed average causal exposure-response function. The simple matching estimator can be regarded as a non-parametric estimator with uniform kernel, thus could be very jagged. To improve smoothness of the curve, we introduce kernel smoothing by either fitting a kernel smoother on w in the matched data to obtain the smoothed average exposure-response function $\hat{\mu}^{(2)}(w)$, or by replacing the uniform kernel in $\hat{\mu}(w)$ with an Epanechnikov/Gaussian kernel to obtain $\hat{\mu}^{(2)}(w)$ as in Jiang et al. (2017).

3.3 Covariate Balance

In this section we introduce two new measures to assess covariate balance in the Design Stage; absolute correlation and (blocked) absolute standardized bias for continuous exposures. The absolute correlation between the exposure and each pre-exposure covariate is a global measure and can

inform whether the whole matched data 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 each exposure level w and is a local measure that informs whether a specific exposure level is balanced or not. 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 $[w_i - \delta, w_i + \delta]$ has exposure W_{ik} and q-dimensional pre-exposure covariates $\mathbf{C}_{ik}$ and outcome Y_{ik}, and 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, the correlations between the exposure and pre-exposure covariates are equal to zero, that is $E[\mathbf{C}^*_{ik}W^*_{ik}] = \mathbf{0}$. We assess 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, 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

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

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. Similarly, 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: $\left|\left|\sum_{i=1}^{I}\sum_{k=1}^{m_i}n_{ik}\mathbf{C}^*_{ik}W^*_{ik}\right|\right|_1/q$.

2) Average absolute standardized bias: $\sum_{i=1}^{I}\left|\left|\frac{\sum_{k=1}^{m_i}\mathbf{C}^*_{ik}}{N}-\frac{\sum_{i'\neq i}\sum_{k=1}^{m_{i'}}\mathbf{C}^*_{i'k}}{N\cdot(I-1)}\right|\right|_1/q$.

3.4 Selecting the parameters (λ, δ)

The scale parameter λ is introduced to control the relative weight that is attributed to the distance measures of the exposure versus the GPS 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 GPS value, resulting in a sacrifice of covariate balance in the matched dataset (Flores et al. 2007). 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 $|| W_j - w ||\leq \delta$.

Both (λ, δ) can be treated as tuning parameters to be selected in the design stage without knowing any outcome information. In finite sample studies, the optimal (λ, δ) could be specified by minimizing a utility function that measures the degree of covariate balance (e.g., the average absolute correlation or the average absolute standardized bias) (McCaffrey et al. 2004, Zhu et al. 2015). Noting that the optimal (λ, δ) aim at achieving covariate balance on the entire matched data, the average absolute correlation would be a suitable measure in practice, moreover, it is more computationally attainable. We summarize our data-driven tuning procedures as follow:

1) Construct the matched data using the matching function with a pair of pre-specified (λ, δ).

2) Calculate the utility function that measures covariate balance on this matched data.

3) Repeat steps 1-2 using grid search on a range of (λ, δ).

4) Find the (λ, δ) which minimize the utility function, leading to the best covariate balance.

The tuning procedures are conducted in the design stage without access to outcome information, thus, this procedure neither biases analyses of outcomes nor requires corrections for multiple inference (Zhu et al. 2015, Rosenbaum 2020).

When applying the proposed approach to data, a caliper δ forbids matches on very different exposure levels, and similarly one can set another caliper on the GPS to improve the quality of the matches but this could lead to more unmatched units (Cochran & Rubin 1973, Rosenbaum 2020). Unmatched units are those who do not have a good representative in the raw observational data. If all 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 such observational data.

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 GPS, or on the estimated GPS, given the fixed scale parameter $\lambda = 1$, the caliper size $\delta = o(N^{-1/3})$ and $N\delta \to \infty$. The summary information is that the proposed matching estimator is asymptotically unbiased, consistent and asymptotically normal with a non-parametric rate $(N\delta)^{-1/2}$ when matching on a scalar covariate (e.g., GPS), yet the properties are not necessary hold if matching on multidimensional covariates, which justifies the GPS matching. Finally, assuming $h \simeq \delta$, the asymptotic normality hold for smoothed estimator with a rate $(Nh)^{-1/2}$.

We begin by defining the conditional means and variances given covariates and given the GPS

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 algebraic expression, we only consider one-to-one nearest neighbor matching on a set of continuous covariates $\mathbf{C}$. All the asymptotic theories can be extended to one-to-M nearest neighbor matching. The matching estimator for $\mu(w)$ can be defined as,

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

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 mean 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 residual. 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* $j_1 = arg\min_{j=1,\ldots,N} ||\mathbf{C}_j - \mathbf{c}||$ *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 A1 hold, if* $\mathbf{C}$ *is scalar, the order of the bias of the proposed matching estimator, that is* $B_\mu(w)$*, is* $O\{(max((N\delta)^{-1}, \delta)\}$*.*

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 $(N\delta)^{-1/2}$ given $\delta = o(N^{-1/3})$, which guarantees the bias does not dominate the asymptotic behaviors of $\hat{\mu}(w)$.

Lemma 4 (Number of replacements) *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 the 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 the 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 the uniform boundedness assumption (A1 in the Supplementary Material) hold. If* $\mathbf{C}$ *is scalar,*

$$\Sigma_1^{-1/2}(N\delta)^{1/2}\{\hat{\mu}(w) - \mu(w)\} \to N\{0,1\}$$
$$\Sigma_1(w) = \frac{1}{N}\sum_{j=1}^{N}\left[\delta K(j)^2\sigma^2_{\mathbf{c}}(W_j, \mathbf{C}_j)I_j(w,\delta)\right]$$

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 GPS matching, we have the following theorem.

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

$$\Sigma_2^{-1/2}(N\delta)^{1/2}\{\hat{\mu}_{\text{GPS}}(w) - \mu(w)\} \to N\{0,1\}$$
$$\Sigma_2(w) = \frac{1}{N}\sum_{j=1}^{N}\left[\delta K(j)^2\sigma^2_{\text{GPS}}\{w, e(w,\mathbf{C}_j)\}I_j(w,\delta)\right]$$

In practice, we never observe the true GPS values, and the GPS has to be estimated prior to matching. Abadie & Imbens (2016) proved and derived 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 GPS is $e(w,\mathbf{c}) = g^{-1}_{\mathbf{\Phi}}(\mathbf{c})$, 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 GPS. We state the following theorem.

Theorem 6 (Asymptotic Normality with estimated GPS) *Assume Assumptions 1-3, the uniform boundedness and the almost sure convergence assumption (A2-3 in the Supplementary Mate-*

rial) hold. The GPS is estimated by a parametric model parameterized by $\mathbf{\Phi}$.

$$\Sigma_2^{-1/2}(N\delta)^{1/2}(\hat{\mu}_{\text{GPS}}(w;\hat{\mathbf{\Phi}}) - \mu(w)) \to N\{0,1\}$$

Theorem 6 states that no matter whether we match on the true GPS or the estimated GPS, the asymptotic properties are unchanged. Importantly, the asymptotic variance remains the same if 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 $o_p\{(N\delta)^{-1/2}\}$, which also holds for many semi-/non-parametric estimations of the GPS, Theorem 6 holds.

Theorem 7 (Asymptotic Normality of smoothed ERF) *Assume Assumptions 1-4, the uniform boundedness and the almost sure convergence assumption (A2-3 in the Supplementary Material) hold. We use a kernel* $K(\cdot)$ *with a symmetric probability density with support* $[-1,1]$ *and bandwidth* h *in* $\hat{\mu}^{(2)}_{\text{GPS}}(w;\hat{\mathbf{\Phi}})$. *Assuming* $h \simeq \delta$ *and a constant* c^*, *we obtain the smoothed estimator* $\hat{\mu}^{(2)}_{\text{GPS}}(w;\hat{\mathbf{\Phi}})$ *which satisfies*

$$\Sigma_2^{-1/2}(Nh)^{1/2}\{\hat{\mu}^{(2)}_{\text{GPS}}(w;\hat{\mathbf{\Phi}}) - \mu(w)\} \to N\{0, c^* \int K^2(u)du\}$$

Theorem 7 is a result of non-parametric statistics, which ensures that both the original and smoothed estimators have an asymptotic normal distribution (Heller 2007, Jiang et al. 2017). Proofs of Theorems 1-7 are provided in the Supplementary Material.

5 Simulations

We conduct extensive simulation studies to evaluate the performance of the proposed matching approach compared to other three state-of-art alternatives, including GPS adjustment estimator (Hirano & Imbens 2004), inverse probability of treatment weighting (IPTW) estimator (Robins et al. 2000), and the non-parametric doubly robust (DR) estimators (Bang & Robins 2005, Kennedy et al. 2017) under various model specifications. For IPTW and DR estimators, we follow common practice of stabilizing and trimming in weighting. We also compare the performance of each estimators when using the parametric linear regression (Hirano & Imbens 2004) and the cross-

validation-based Super Learner algorithm (Van der Laan et al. 2007, Kennedy et al. 2017) to estimate the GPS.

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 six specifications of the GPS model,

1) $W = 9\{-0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)\boldsymbol{C}\} + 17 + N(0, 5)$

2) $W = 15\{-0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)\boldsymbol{C}\} + 22 + T(2)$

3) $W = 9\{-0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)\boldsymbol{C}\} + 3/2C_3^2 + 15 + N(0, 5)$

4) $W = \frac{49\exp(\{-0.8+(0.1,0.1,-0.1,0.2,0.1,0.1)\boldsymbol{C}\})}{1+\exp(\{-0.8+(0.1,0.1,-0.1,0.2,0.1,0.1)\boldsymbol{C}\})} - 6 + N(0, 5)$

5) $W = \frac{42}{1+\exp(\{-0.8+(0.1,0.1,-0.1,0.2,0.1,0.1)\boldsymbol{C}\})} - 18 + N(0, 5)$

6) $W = 7\log(\{-0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)\boldsymbol{C}\}) + 13 + N(0, 4)$

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,

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

For these six specifications we vary the sample size $N(= 200, 1000, 5000)$ resulting in a total of eighteen scenarios. For each scenario, we generate 500 datasets.

After generating the data we estimate the exposure-response function for each simulation scenario using four different approaches including the proposed matching approach and three state-of-art alternatives. For IPTW and DR estimators, we follow common practice of stablizing and

trimming in weighting (see the Supplementary Material for details). When estimating the GPS, we use both parametric linear regression assuming normal error (Hirano & Imbens 2004) and the cross-validation-based Super Learner algorithm (Van der Laan et al. 2007, Kennedy et al. 2017).

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:

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

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

where $\hat{\mathcal{W}}^*$ denotes a restricted version of the support of $\hat{\mathcal{W}}$, excluding 10% of mass at the boundaries to avoid boundary instability.

5.2 Balance Assessment

The matching framework, which maintains intact the unit of analysis, provides a transparent way to assess covariate balance. In practice, we can compare the values of covariates balance measures, for example absolute correlations, described in Section 3.3 between the matched dataset and unadjusted observational data. If the absolute correlations for each of the observed covariates in the matched dataset are substantially smaller than those in unadjusted observational data, we conclude 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 results from simulation settings where we vary the six specifications of the GPS model using Super Learner under sample size $N = 5000$. We assess balance by calculating absolute correlation for each of six covariates. We see that balance improves substantially across all six covariates for all six simulation settings. Under five out of six setting (scenario $1, 3-6$), absolute correlations for all confounders are < 0.10, which indicates excellent balance for observed

covariates within the matched dataset. Under settings of extreme values for the GPS (scenario 2), the proposed matching procedure largely improves balance for all covariates, though there is still evidence of imbalance.

5.3 Simulation results

Table 1 shows the simulation results for the settings where we estimate the GPS model using Super Learner algorithms, while Table 2, shows results where we estimate the GPS model using parametric linear regressions. We found similar results across the four implementation approaches regardless of how the GPS model was estimated (Super Learner algorithm vs. parametric regression). More specifically, when the GPS model is correctly specified and does not contain extreme values (scenario 1), all approaches perform reasonably well. The proposed matching and Kennedy's doubly robust approaches, in general, outperform the standard IPTW and GPS adjustment approaches, in terms of both absolute bias and MSE. The matching estimator provides the smallest absolute bias, yet Kennedy's doubly robust estimator provides smaller MSE.

When the GPS model is correctly specified yet includes extreme GPS values (scenario 2), the GPS adjustment, IPTW and doubly robust estimators all produce very large MSE, and are not able to reduce confounding bias even as the sample sizes increase. There is plenty of literature suggesting stabilizing weights and trimming extreme weights under binary/categorical exposure settings Harder et al. (2010), Crump et al. (2009), Yang et al. (2016), yet the guidelines on handling extreme weights under continuous exposure regimes are sparse. In these simulation studies, we found that the common practical guidelines for trimming (cap the stabilized weight at 10 (Harder et al. 2010)) does not provide remedy. In contrast, our matching estimator is robust to the extreme GPS values, creating much more stable estimation than any of the other state-of-art alternatives evaluated. Importantly, the absolute bias and MSE of the proposed matching estimator decreases as the sample sizes increases.

When the GPS model is misspecified in various ways (scenarios 3-6), the proposed matching approach consistently provides stable small bias reduction and MSE no matter how the GPS were estimated. It is worth noting that when the GPS is modelled by parametric linear regressions (which is a common approach), matching provided notable better performances compared to all

other approaches. This finding is aligned with results from Waernbaum (2012) under binary exposure settings, showing that when matching on a parametric model (e.g., a propensity score), the matching estimator is robust to model misspecifications if the misspecified model belongs to the class of covariate scores. This implies there are multiple possibilities for the matching estimator to make reliable inference, which highlights the robustness of the method. When the GPS is modelled by a Super Learner, the performances of other approaches improves, likely because the flexible modelling techniques can effectively recover the correct form of the GPS. Yet still matching outperforms standard IPTW and GPS adjustment approaches both in terms of absolute bias and MSE, though it is slightly less efficient than the DR estimator.

In general, via a comprehensive set of simulations, we see that the proposed matching approach consistently performs well including under settings in presence of extreme estimated GPS values and/or of GPS model misspecifications. This robust 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 GPS model compared to GPS adjustment and weighting-type approaches.

6 Data Application

In air pollution epidemiology studies (e.g., Pope III et al. (2019)), the scientific question commonly proposed is "whether and in what magnitude the exposure is (causally) associated with the adverse health outcome?". The unit of analysis can often be individual, zip code, county, and so on. Researchers collect information about exposure, outcome and a set of characteristic covariates linked to each unit. We apply the proposed matching method to estimate the effect of long-term $PM_{2.5}$ exposure on all-cause mortality. To this end, we use the largest-to-date Medicare enrollee cohort across the contiguous US from 2000 to 2016. This study population includes a total of 68.5 million individuals, who reside in $31,414$ zip codes. We construct counts corresponding to the all-cause mortality for Medicare enrollees for each zip code per year across the US. daily $PM_{2.5}$ exposures were estimated at a 1km $\times$ 1km grid cell resolution using a spatio-temporal prediction model with excellent predictive accuracy (cross validation $R^2 = 0.86$)(Di et al. 2019). To obtain annual average $PM_{2.5}$ at each zip code, we aggregate the gridded concentrations using area-weighted

averages. We assign the annual average $PM_{2.5}$ to individuals who reside in that zip code for each calendar year. The range of predicted annual average $PM_{2.5}$ from 2000 to 2016 was 0.01 – 30.92 $\mu g/m^3$ with 1% and 99% quantiles equal to $(2.76, 17.16)$.

Design Stage. We estimate the GPS by using gradient boosting machine (Chen & Guestrin 2016, Zhu et al. 2015), with annual $PM_{2.5}$ exposure as the outcome and 19 potential confounders, including population demographic information, Medicaid information, meteorological information and spatio-temporal trends (census region/year). After obtaining the estimated GPS, we implement our proposed matching procedure. Specifically, we use a pre-specified two-dimensional Manhattan matching function with scale parameter $\lambda = 1$ and caliper $\delta = 0.16$ (creating 100 equidistant levels throughout the whole exposure range). The choice of (λ, δ) follows the data-driven tuning procedures in Section 3.4. We construct the matched dataset by collecting all imputed observations.

We assess covariate balance by calculating the absolute correlation for each potential confounders as discussed in Section 3.3. The GPS matching implementation largely improves covariate balance for 16 out of 19 potential confounders. The average absolute correlation is 0.19 before matching, whereas, the average absolute correlation is 0.04 after matching (See Figure 2). Importantly, although time trend (year) has a strong imbalance before matching, it is balanced after matching.

Analysis Stage. After obtaining the matched set, we fit the kernel smoothing on the data to estimate the causal exposure-response function relating long-term $PM_{2.5}$ levels to all-cause mortality rate. We construct the point-wise Wald 95% confidence band for the exposure-response function using m-out-of-n bootstrap, a modified bootstrap method (Bickel et al. 2012). To avoid potential ill-behaviours at the support boundaries, consistent with Liu et al. (2019), Di et al. (2017), we exclude the highest 1% and lowest 1% of $PM_{2.5}$ exposures.

Figure 3 shows the average causal exposure-response function. To our knowledge, this is the first exposure-response curve assessing the effects of long-term $PM_{2.5}$ on all-cause mortality using a causal inference approach to account for measured confounders, which provides the most robust evidence to date of the causal link in environmental epidemiology. We find an consistently harmful causal relationship between mortality and long-term $PM_{2.5}$ exposures across the range of annual average $PM_{2.5}$ (2.76–17.16 $\mu g/m^3$) in the entire Medicare enrollees across the continental US from

2000 to 2016. There is currently 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 reporting a strong association between long-term exposure to $PM_{2.5}$ and adverse health outcomes at low exposure levels. Importantly, the curve is steeper at exposure levels lower than the current standards, indicating aggravated harmful effects at exposure levels even below the national standards. By implementing a univariate Poisson regression on the matched set, we find each 10 $\mu g/m^3$ increase of exposure level of annual average $PM_{2.5}$ causes an approximately 7.0% increase in all-cause mortality rate.

7 Discussion

We developed an innovative causal inference approach using enormous observational data in settings with continuous exposures, and introduced a new framework for GPS caliper matching. Our proposed approach fills an important gap in the literature as it provides a theoretically-justified generalization 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). Specifically, it 1) is robust to misspecification of the GPS model, especially in the presence of extreme values (Kang et al. 2007, Waernbaum 2012), 2) allows flexible specifications of the outcome model, including non-parametric Kernel methods or flexible machine learning approaches, 3) maintains intact the unit of analysis and, thus, allows for the transparent assessment of covariate balance, 4) provides a hot deck imputation which can be straightforwardly extended to estimate other distributional causal estimands, e.g. quantile exposure effects (Andridge & Little 2010, Yang & Zhang 2020). In addition, we introduced new measures to evaluate covariate balance, and described the way to assess balance based on these measures.

The development of our matching approach is motivated by the application on the evaluation of causal effects of long term exposure to $PM_{2.5}$ on the risk of all-cause mortality. In health policy there

is increased scrutiny on the strength of evidence for health effects of particulate matter exposure, with emphasis on the robustness of the methodology and interpertability (Goldman & Dominici 2019). Matching approaches are advantageous as they are robust to model misspecification and allow for straightforward outcome/sensitivity analysis strengthening the interpertability of the results. We applied the proposed matching approach to estimate causal effects between long-term $PM_{2.5}$ and all-cause mortality on a massive Medicare administrative data cohort. We found a positive and near-linear causal exposure-response relationship between long-term $PM_{2.5}$ and all cause mortality among the entire US Medicare enrollees (2000-2016). The slope of the causal exposure-response function is steeper when the $PM_{2.5}$ concentrations are below 6 $\mu g/m^3$, which is consistent with previous findings based on observational studies (Di et al. 2017, Liu et al. 2019). Our estimates were obtained under a robust causal inference framework which allows for transparent assessment of covariate balance and helps unveil evidence of causality. Such finding suggests that even countries, such as the United States, with relatively good air quality could still see public health benefits from further reduction of ambient $PM_{2.5}$ concentrations (i.e., below the current standards) (Balmes 2019).

Some air pollution studies have been conducted using propensity score-based analyses, however, applied researchers often dichotomize or categorize continuous exposure variables in order to utilize propensity score methods (Baccini et al. 2017, Wu et al. 2019). The GPS caliper matching approach introduced in this paper is the first matching approach which allows for the estimation of causal effects on continuous exposures and the assessment of covariate balance in a transparent way. Computational feasibility is another urgent consideration during the method development for such large-scale studies. The proposed matching with replacement eases the computational burden (Imbens & Rubin 2015), and has the capability of utilizing parallel computing to accelerate the matching procedure.

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) starts to relax it by allowing interference, yet future work needs to be done by combining 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. 2019) and trimming techniques to ensure overlap (Crump et al. 2009, Harder et al. 2010, Yang et al. 2016). However, under continuous exposure cases, since the overlap is defined by a density function on a Lebesgue set, it is conceptually hard to check it directly via finite samples. One potential way is to 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. 2019), yet no current approach is able to directly verify the overlap on continuous scale. Future work is needed to develop rigorous approaches to check/ensure overlap in continuous exposure settings.

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 counter-factual outcome. The limitation of weak unconfoundedness is that it only allows for the identification and estimation of the population average causal exposure-response function, but not for the identification of the average causal exposure-response function for specific subpopulations. However, it is valid when the interest is in estimating the causal effects across the whole population and not subpopulations. In addition, as with other (generalized) propensity score-based analyses, this approach does not resolve the potential for bias due to unmeasured confounding, in which case the unconfoundedness assumption is violated. By choosing a suitable degree of approximation, i.e., choosing δ, under the local weak unconfoundedness assumption, we identify the theoretical point at which the proposed matching estimator achieves desirable asymptotic properties. The smoothness assumption is essentially the standard smoothness condition imposed in non-parametric regression problems. Also, we require the rate of smoothness, i.e., the bandwidth, to satisfy $h \simeq \delta$, to ensure the bias from matching discrepancy is asymptotically negligible. In finite-sample practice, the (δ, h) are considered as tuning parameters and searched via observational data by cross-validation. The focus of this paper is not to find a non-parametric estimator with the sharpest rate of convergence; thus, we obtain the asymptotically unbiased matching estimator via under-smoothing. A natural extension is to generalize the bias-corrected matching estimator proposed in Abadie & Imbens (2011) into our non-parametric settings, which has the potential to obtain sharper results on the rate of convergence.

Finally, we believe the simplicity and generality of our matching framework has the potential of promoting awareness of causal inference in future science and policy-relevant research, especially in fields where interventions (or named exposures/treatments) are naturally continuous such as in environmental research.

Acknowledgement

The authors are grateful to Junwei Lu, Ziyang Wei, Agnese Panzera, Jose 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.

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. (2011), 'Bias-corrected matching estimators for average treatment effects', *Journal of Business & Economic Statistics* **29**(1), 1–11.

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

Andridge, R. R. & Little, R. J. (2010), 'A review of hot deck imputation for survey non-response', *International statistical review* **78**(1), 40–64.

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.

Balmes, J. R. (2019), 'Do we really need another time-series study of the pm2. 5–mortality association?'.

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

Bickel, P. J., Götze, F. & van Zwet, W. R. (2012), Resampling fewer than n observations: gains, losses, and remedies for losses, *in* 'Selected works of Willem van Zwet', Springer, pp. 267–297.

Bind, M.-A. (2019), 'Causal modeling in environmental health', *Annual review of public health* **40**, 23–43.

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.

Carone, M., Dominici, F. & Sheppard, L. (2019), 'In pursuit of evidence in air pollution epidemiology: The role of causally driven data science', *Epidemiology* .

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining', ACM, pp. 785–794.

Cochran, W. G. & Rubin, D. B. (1973), 'Controlling bias in observational studies: A review', *Sankhyā: The Indian Journal of Statistics, Series A* pp. 417–446.

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., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A. et al. (2019), 'An ensemble-based model of pm2. 5 concentration across the contiguous united states with high spatiotemporal resolution', *Environment international* **130**, 104909.

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.

Dominici, F., Greenstone, M. & Sunstein, C. R. (2014), 'Particulate matter matters', *Science* **344**(6181), 257–259.

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.

Goldman, G. T. & Dominici, F. (2019), 'Don't abandon evidence and process on air pollution policy', *Science* **363**(6434), 1398–1400.

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.

Heller, G. (2007), 'Smoothed rank regression with censored data', *Journal of the American Statistical Association* **102**(478), 552–559.

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., Kim, I. S. & Wang, E. (2018), 'Matching methods for causal inference with time-series cross-section data', *Princeton University* **1**.

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.

Jiang, R., Lu, W., Song, R. & Davidian, M. (2017), 'On estimation of optimal treatment regimes for maximizing t-year survival probability', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(4), 1165–1185.

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.

Liu, C., Chen, R., Sera, F., Vicedo-Cabrera, A. M., Guo, Y., Tong, S., Coelho, M. S., Saldiva, P. H., Lavigne, E., Matus, P. et al. (2019), 'Ambient particulate air pollution and daily mortality in 652 cities', *New England Journal of Medicine* **381**(8), 705–715.

Lu, B., Greevy, R., Xu, X. & Beck, C. (2011), 'Optimal nonbipartite matching and its statistical applications', *The American Statistician* **65**(1), 21–30.

McCaffrey, D. F., Ridgeway, G. & Morral, A. R. (2004), 'Propensity score estimation with boosted regression for evaluating causal effects in observational studies.', *Psychological methods* **9**(4), 403.

Peters, A., Künzli, N., Forastiere, F. & Hoffmann, B. (2019), 'Promoting clean air: combating fake news and denial', *The Lancet Respiratory Medicine* **7**(8), 650–652.

Pope III, C. A., Coleman, N., Pond, Z. A. & Burnett, R. T. (2019), 'Fine particulate air pollution and human mortality: 25+ years of cohort studies', *Environmental Research* p. 108924.

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. (2020), 'Modern algorithms for matching in observational studies', *Annual Review of Statistics and Its Application* **7**.

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.

Stuart, E. A. (2010), 'Matching methods for causal inference: A review and a look forward', *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**(1), 1.

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.

Van der Laan, M. J., Polley, E. C. & Hubbard, A. E. (2007), 'Super learner', *Statistical applications in genetics and molecular biology* **6**(1).

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, M., Aaron, C. P., Madrigano, J., Hoffman, E. A., Angelini, E., Yang, J., Laine, A., Vetterli, T. M., Kinney, P. L., Sampson, P. D. et al. (2019), 'Association between long-term exposure to ambient air pollution and change in quantitatively assessed emphysema and lung function', *Jama* **322**(6), 546–556.

Wu, X., Braun, D., Kioumourtzoglou, M.-A., Choirat, C., Di, Q., Dominici, F. et al. (2019), 'Causal inference in the context of an error prone exposure: air pollution and mortality', *The Annals of Applied Statistics* **13**(1), 520–547.

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.

Yang, S. & Zhang, Y. (2020), 'Double score matching estimators of average and quantile treatment effects', *arXiv preprint arXiv:2001.06049* .

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.

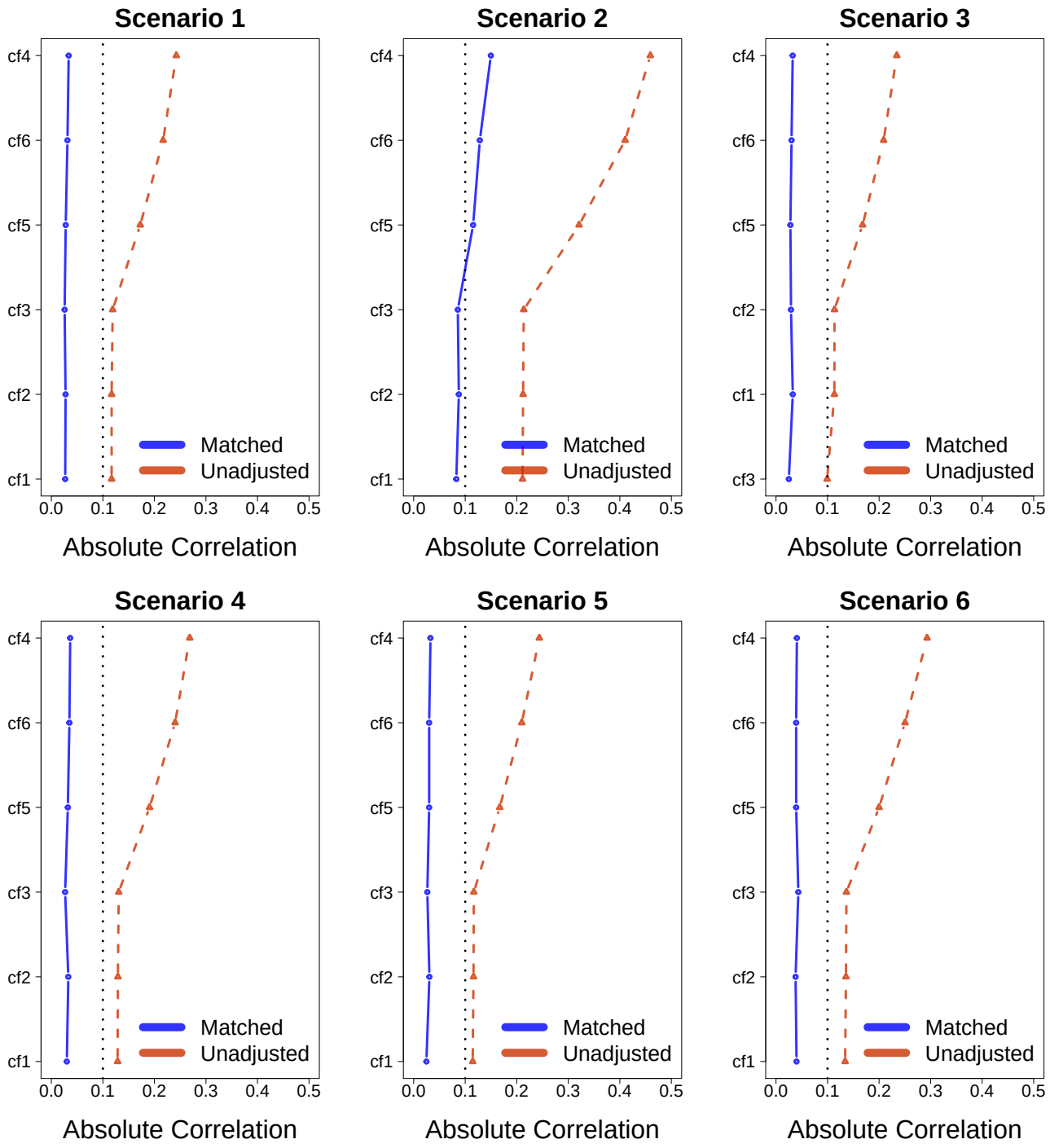

Figure 1: Absolute Correlations (ACs). Each panel represents the ACs for each covariate in the matched dataset (solid line) and original dataset (dashed line) under six simulation settings where GPS model specifications vary. The dotted line represents the cut-off of covariate balance suggested by Zhu et al. (2015). The GPS was estimated by using Super Learner under sample size $N = 5000$. GPS matching improves covariate balance for all six covariates in all settings.

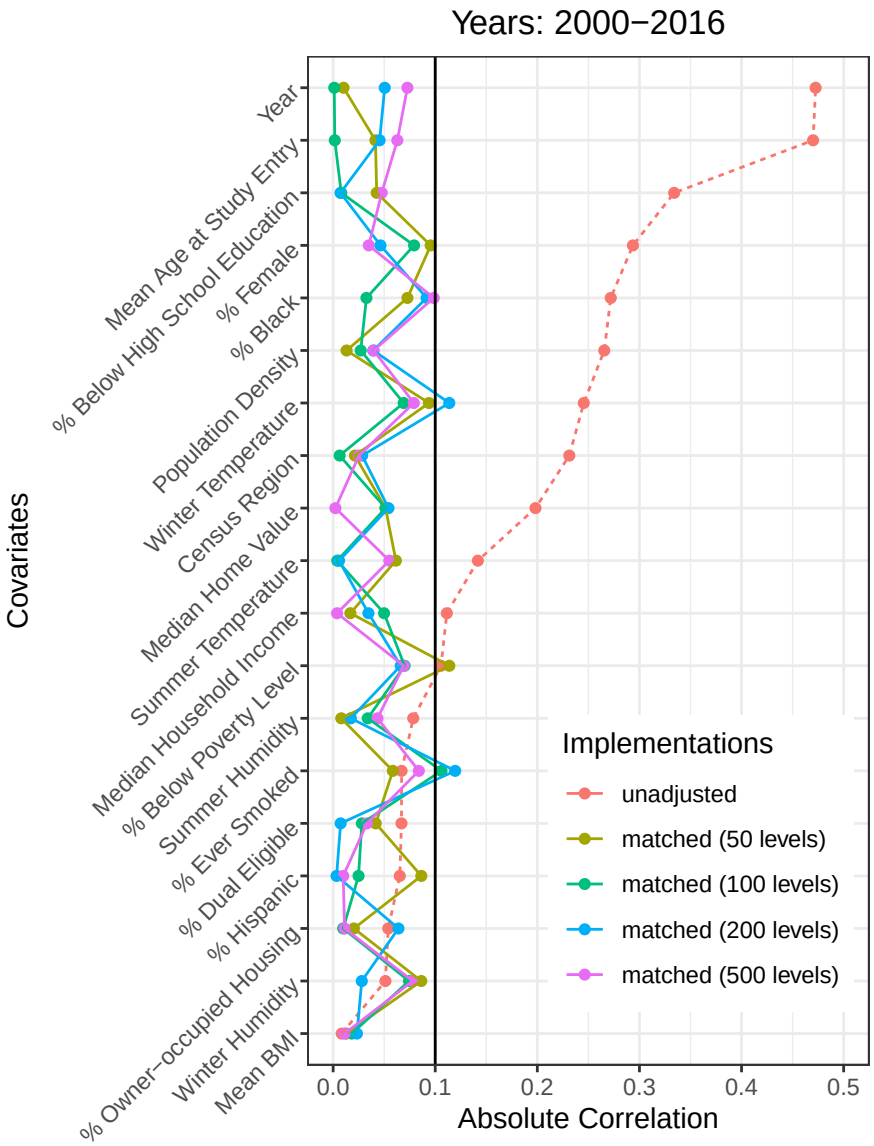

Figure 2: Absolute Correlations (ACs). The figure represents the ACs for each covariate in the matched dataset (solid line) and original dataset (dashed line). For matched dataset, we consider four scenarios varying the numbers of blocks (which is equivalent to varying caliper sizes). The dotted line represents the cut-off of covariate balance suggested by Zhu et al. (2015). In general, GPS matching substantially improves covariate balance for these potential confounders. The average AC is 0.19 before matching, and 0.04 after matching for all scenarios, in particular the average AC is minimized when we use 100 levels (i.e., caliper $\delta = 0.16$). Importantly, time trend (year) has a strong imbalance before matching, yet is balanced after matching.

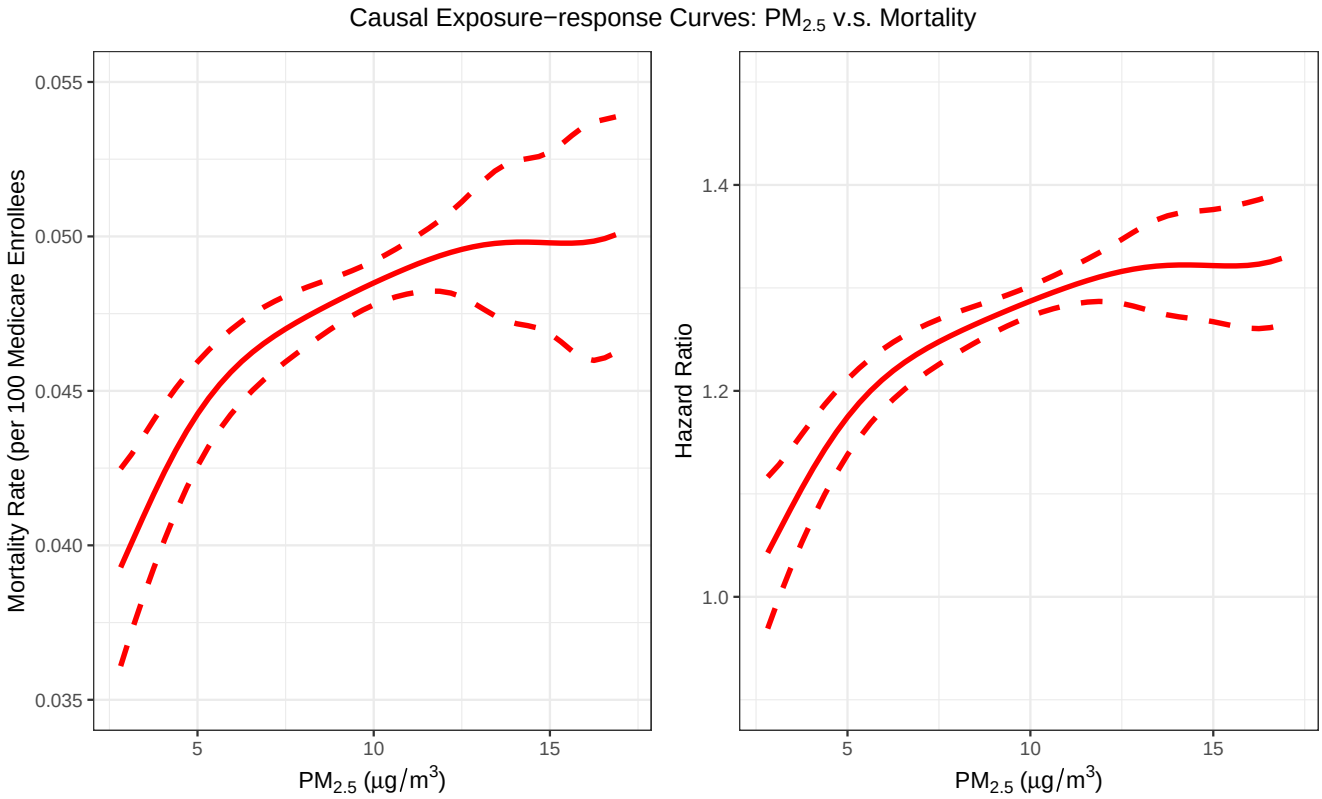

Figure 3: The causal exposure-response function relating all-cause mortality to long-term $PM_{2.5}$ exposure. The left panel presents the smoothed causal exposure-response curve in mortality rate obtained by kernel smoothing with optimal bandwidth (red solid line) and its point-wise confidence band calculated by m-out-of-n bootstrap. The right panel is the smoothed curve in hazard ratio with its point-wise confidence band. The GPS was estimated by using gradient boosting machine (Chen & Guestrin 2016).

Table 1: Absolute Bias and Mean Squared Error (MSE). We estimate the GPS using Super Learner models, based on 500 simulation replicates

GPS generations	N	**Matching**	Adjustment	IPTW	DR	IPTW (trim)	DR (trim)
1) $N(0,5)$-distributed residuals	200	0.83 (4.42)	1.49 (5.72)	2.11 (3.95)	0.95 (2.76)	2.11 (3.94)	1.20 (3.31)
	1000	0.40 (2.05)	1.33 (2.93)	2.04 (3.48)	0.73 (1.87)	2.04 (3.48)	0.49 (1.47)
	5000	0.12 (1.36)	0.97 (1.93)	1.76 (3.51)	0.54 (1.22)	1.77 (3.51)	0.37 (0.80)
2) $t(2)$-distributed residuals	200	2.98 (7.40)	8.63 (73.44)	4.65 (24.61)	17.66 (181.06)	4.66 (24.71)	17.79 (180.40)
	1000	2.15 (4.38)	11.17 (172.47)	3.11 (9.42)	16.82 (135.30)	3.14 (9.51)	36.49 (458.34)
	5000	1.55 (2.76)	33.04 (157.05)	4.54 (9.81)	25.14 (155.74)	4.63 (10.23)	46.02 (245.88)
3) 2nd order term	200	1.24 (4.81)	1.57 (7.26)	2.40 (4.40)	1.00 (2.64)	2.40 (4.39)	1.38 (3.25)
	1000	0.64 (2.28)	1.67 (3.67)	2.17 (3.97)	0.85 (1.87)	2.17 (3.97)	0.68 (1.58)
	5000	0.29 (1.48)	1.26 (2.27)	1.91 (3.97)	0.65 (1.54)	1.91 (3.97)	0.65 (1.22)
4) logistic link	200	1.25 (4.85)	1.92 (6.75)	2.22 (4.29)	0.91 (2.71)	2.22 (4.29)	1.17 (3.41)
	1000	0.63 (2.17)	1.69 (3.23)	2.05 (3.69)	0.67 (1.63)	2.05 (3.69)	0.54 (1.45)
	5000	0.25 (1.43)	1.17 (2.15)	1.82 (3.76)	0.60 (1.27)	1.82 (3.76)	0.44 (0.95)
5) 1-logistic link	200	0.51 (4.43)	1.24 (4.59)	1.44 (3.08)	0.62 (2.37)	1.44 (3.08)	0.94 (2.87)
	1000	0.47 (2.02)	1.62 (2.70)	1.63 (2.77)	0.68 (1.66)	1.63 (2.77)	0.43 (1.35)
	5000	0.14 (1.33)	0.99 (1.71)	1.27 (2.55)	0.51 (1.13)	1.27 (2.55)	0.29 (0.77)
6) log link	200	1.41 (4.86)	2.76 (11.79)	2.88 (5.43)	1.72 (6.54)	2.88 (5.43)	1.58 (6.48)
	1000	1.42 (2.92)	2.36 (9.12)	2.54 (3.99)	0.83 (2.42)	2.54 (3.99)	0.66 (2.21)
	5000	1.27 (2.74)	5.67 (17.03)	3.02 (4.15)	0.72 (1.69)	3.02 (4.15)	0.70 (1.65)

Notes: Matching = the proposed GPS caliper matching; Adjustment = includes GPS as covariates in a outcome model proposed in Hirano & Imbens (2004); IPTW = inverse probability of treatment weighted; DR = doubly robust proposed in Kennedy et al. (2017); trim = trim the stabilized weight that larger than 10.

Table 2: Absolute Bias and Mean Squared Error (MSE). We estimate the GPS using standard linear regression models, based on 500 simulation replicates

GPS generations	N	**Matching**	Adjustment	IPTW	DR	IPTW (trim)	DR (trim)
1) $N(0,5)$-distributed residuals	200	1.02 (3.87)	1.19 (3.50)	2.17 (4.08)	1.08 (3.71)	2.17 (4.08)	1.22 (3.29)
	1000	0.54 (1.90)	1.31 (2.49)	1.97 (3.59)	0.90 (2.31)	1.98 (3.56)	0.61 (1.47)
	5000	0.21 (1.29)	1.01 (1.90)	1.42 (3.23)	0.60 (1.45)	1.44 (3.16)	0.49 (0.87)
2) $t(2)$-distributed residuals	200	3.16 (6.98)	3.45 (42.78)	* (*)	* (*)	5.69 (22.27)	14.61 (108.74)
	1000	2.20 (4.17)	* (*)	85.08 (*)	* (*)	7.53 (21.46)	48.29 (704.54)
	5000	1.44 (2.91)	* (*)	* (*)	* (*)	* (*)	114.06 (700.18)
3) 2nd order term	200	1.41 (4.28)	1.81 (4.25)	2.58 (4.80)	2.18 (4.85)	2.59 (4.73)	1.60 (3.52)
	1000	0.86 (2.07)	1.50 (2.71)	2.00 (4.10)	2.35 (4.73)	2.07 (3.94)	0.85 (1.70)
	5000	0.55 (1.42)	1.16 (2.07)	1.51 (4.01)	2.68 (13.77)	1.55 (3.44)	0.83 (1.33)
4) logistic link	200	1.35 (4.27)	1.61 (4.13)	2.46 (4.58)	1.29 (3.59)	2.47 (4.57)	1.29 (3.39)
	1000	0.62 (2.03)	1.71 (2.90)	1.98 (3.89)	0.89 (2.56)	2.01 (3.83)	0.63 (1.49)
	5000	0.34 (1.36)	1.19 (2.11)	1.44 (3.56)	0.63 (1.49)	1.46 (3.46)	0.56 (1.01)
5) 1-logistic link	200	0.60 (3.81)	1.14 (3.09)	1.30 (3.19)	0.92 (3.55)	1.31 (3.17)	0.53 (2.71)
	1000	0.43 (1.84)	1.48 (2.32)	1.51 (2.77)	0.83 (2.20)	1.50 (2.72)	0.35 (1.32)
	5000	0.19 (1.24)	1.00 (1.63)	1.05 (2.47)	0.54 (1.26)	1.04 (2.37)	0.22 (0.75)
6) log link	200	1.26 (4.17)	3.30 (9.87)	2.68 (5.39)	2.62 (45.00)	2.74 (5.26)	0.99 (4.18)
	1000	0.97 (2.17)	2.46 (4.18)	2.51 (4.09)	3.57 (97.76)	2.53 (4.07)	0.44 (1.80)
	5000	0.62 (1.48)	2.55 (3.59)	4.06 (47.33)	10.51 (146.82)	1.97 (4.96)	0.91 (1.42)

Notes: Matching = the proposed GPS caliper matching; Adjustment = includes GPS as covariates in a outcome model proposed in Hirano & Imbens (2004); IPTW = inverse probability of treatment weighted; DR = doubly robust proposed in Kennedy et al. (2017); trim = trim the stabilized weight that larger than 10. * represented values larger than 1000 or more than 50% of simulations fail to converge.