CAUSAL INFERENCE IN THE CONTEXT OF AN ERROR PRONE EXPOSURE: AIR POLLUTION AND MORTALITY[1]

BY XIAO WU*, DANIELLE BRAUN*,
MARIANTHI-ANNA KIOUMOURTZOGLOU[†], CHRISTINE CHOIRAT*,
QIAN DI* AND FRANCESCA DOMINICI*

Harvard T.H. Chan School of Public Health and
Columbia University Mailman School of Public Health[†]

We propose a new approach for estimating causal effects when the exposure is measured with error and confounding adjustment is performed via a generalized propensity score (GPS). Using validation data, we propose a regression calibration (RC)-based adjustment for a continuous error-prone exposure combined with GPS to adjust for confounding (RC-GPS). The outcome analysis is conducted after transforming the corrected continuous exposure into a categorical exposure. We consider confounding adjustment in the context of GPS subclassification, inverse probability treatment weighting (IPTW) and matching. In simulations with varying degrees of exposure error and confounding bias, RC-GPS eliminates bias from exposure error and confounding compared to standard approaches that rely on the error-prone exposure. We applied RC-GPS to a rich data platform to estimate the causal effect of long-term exposure to fine particles (PM$_{2.5}$) on mortality in New England for the period from 2000 to 2012. The main study consists of 2202 zip codes covered by 217,660 1 km × 1 km grid cells with yearly mortality rates, yearly PM$_{2.5}$ averages estimated from a spatio-temporal model (error-prone exposure) and several potential confounders. The internal validation study includes a subset of 83 1 km × 1 km grid cells within 75 zip codes from the main study with error-free yearly PM$_{2.5}$ exposures obtained from monitor stations. Under assumptions of noninterference and weak unconfoundedness, using matching we found that exposure to moderate levels of PM$_{2.5}$ ($8 < $ PM$_{2.5} \leq 10$ μg/m^3) causes a 2.8% (95% CI: 0.6%, 3.6%) increase in all-cause mortality compared to low exposure (PM$_{2.5} \leq 8$ μg/m^3).

1. Introduction. When trying to estimate exposure effects, observational studies are widely used but are susceptible to some well-recognized sources of bias, including but not limited to (1) exposure measurement error and (2) confounding. The measurement error can arise from using mismeasured exposures in

the analysis, since obtaining estimates of the error-free exposures is not always feasible. In addition, there is a confounding problem due to the lack of randomization in observational studies.

Measurement error approaches have been extensively studied in regression problems (Fuller (2009)). There is a large literature on this topic both in linear and nonlinear regression models, including likelihood-based approaches, regression calibration, simulation extrapolation (SIMEX) and Bayesian approaches (Carroll et al. (2006)). In addition, in the context of air pollution, which is the motivation of this work (see Section 4), methods to adjust for measurement error under a noncausal framework have been previously proposed; Alexeeff, Carroll and Coull (2016), Dominici, Zeger and Samet (2000), Gryparis et al. (2009), Hart et al. (2015), Szpiro, Sheppard and Lumley (2011), Van Roosbroeck et al. (2008). Many of them consider a "widely used, effective [and] reasonably well-investigated" (Pierce and Kellerer (2004)) method to adjust for measurement error, namely regression calibration. This method utilizes the following combined study designs: a large main study, for which W (the error-prone exposure) and $\mathbf{D}$ (a set of error-free covariates) are observed, and a smaller validation study, for which in addition to observing $(W, \mathbf{D})$, X is also observed (the error-free exposure). The basic idea of regression calibration is to fit the regression model $X|W, \mathbf{D}$ in the validation study, and use the coefficients from this model to predict X in the main study. After this prediction step, the proposed statistical analysis is performed on the main study with the predicted error-free exposures, $\hat{X}$, to obtain parameter estimates, and either bootstrap or the sandwich variance estimation are used to obtain adjusted standard errors. The simplicity of this algorithm disguises its power.

In addition, observational studies are susceptible to confounding bias by factors that are associated with both the exposure and outcome of interest. Failure to account for them in the analysis may lead to substantial bias. Although most studies adjust for confounding, many do so by simply including the potential confounders as covariates in the outcome model. However, doing so may lead to model misspecification and allows for residual confounding (Rothman, Greenland and Lash (2008)). Therefore, addressing confounding bias in a causal inference framework can be advantageous. A common approach for confounding adjustment in this framework is using propensity scores, the probability of a unit being assigned to a particular treatment, or exposure in our setting, given the pretreatment confounders.

Using propensity scores to adjust for confounding in a causal inference framework was first introduced by Rosenbaum and 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—for propensity score estimation see Dehejia and Wahba (1998, 1999), McCaffrey, Ridgeway and Morral (2004); for propensity score implementation see Harder, Stuart and Anthony (2010), Hirano, Imbens and Ridder (2003), Robins, Hernán and Brumback (2000), Rosenbaum and Rubin (1984). A common technique for

estimation of the propensity score is by fitting a logistic regression model to predict the treatment (or in our setting exposure) with potential confounders included as predictors in the model. Three common techniques for propensity score implementation are matching, inverse probability of treatment weighting (IPTW) and subclassfication (Harder, Stuart and Anthony (2010)). This traditional propensity score framework is only able to handle binary exposures.

In some cases the interest is in estimating the exposure effect for a categorical exposure. To handle categorical exposures, a generalized propensity score (GPS) framework has been developed (Imbens (2000)). Imbens (2000) developed a natural analogue to propensity score estimation under categorical exposures, which uses multinomial logistic regression, instead of logistic regression, to predict multiple exposure categories with all potential confounders included as predictors. They describe an analogue to IPTW for categorical exposures. Although there is no natural analogue to matching and subclassification for the GPS (Imbens (2000), Lechner (2001), Rassen et al. (2013)), Yang et al. (2016) propose an alternative way to estimate causal effects using matching and subclassification for categorical exposures by averaging potential outcomes separately for each of the exposure categories.

Measurement error adjustment for binary exposures under a causal inference framework has been studied by Babanezhad, Vansteelandt and Goetghebeur (2010) and Braun et al. (2017). In Babanezhad, Vansteelandt and Goetghebeur (2010), authors investigate how mismeasured exposures impact the estimation of the causal effects using four different approaches to adjust for confounding, including ordinary least squares (OLS), IPTW, G-estimation and propensity score covariate adjustment. They derive the asymptotic bias for these four estimators, and show they are equally affected by measurement error under linear models when exposure measurement error is independent of the confounders, but not otherwise. Braun et al. (2017) proposes a two-step maximum likelihood approach using validation data to adjust for the measurement error, which effectively corrects for measurement error in binary exposures under a causal inference framework. Specifically, they first use a likelihood based adjustment to correct for measurement error in the propensity score model and estimate an adjusted propensity score. Next, based on the adjusted propensity score, they perform a likelihood-based adjustment on the outcome model to adjust for measurement error in the exposure variable directly. These approaches, however, assume binary exposures and are not directly applicable to a categorical exposure.

In this work, we focus on settings for which we have a continuous exposure measured with error, yet our interest is in estimating causal effects on a categorical scale. We propose a regression calibration (RC)-based adjustment to adjust for the measurement error in the exposure combined with GPS to adjust for confounding (RC-GPS). The RC model is fitted using the continuous exposure, regressing the true exposure on the error-prone exposure and additional covariates on which the

measurement error could depend. Outcome analysis adjusting for confounding using GPS is then conducted after transforming the corrected continuous exposure into a categorical exposure. The proposed method is innovative in the following ways: (1) it provides a correction for measurement error in the exposure for both design and analysis phases with GPS; (2) GPS implementations can be paired with any GLM outcome model (e.g., log-linear model); (3) we show how standardized bias can be used to assess balance in the context of GPS analysis for categorical exposures.

In Section 2 we introduce the proposed adjustment. We then run extensive simulations to assess the performance of our proposed adjustment in Section 3. We apply our proposed approach to investigate the effect between long-term $PM_{2.5}$ exposure and mortality in New England (VT, NH, CT, MA, RI and ME), using zip code aggregated data from Medicare. For the entire Medicare population (main study), long-term exposure to fine particles ($PM_{2.5}$) is determined from a spatio-temporal model that uses multiple different sources as input (meteorological, land use variables, satellite data, etc.). $PM_{2.5}$ exposure based on these predictions is inaccurate, but for a subset of zip codes (validation study) we have actual $PM_{2.5}$ concentrations measured at monitors (error-free exposure). Although in reality $PM_{2.5}$ concentrations measured at monitors could still contain measurement error, for example, instrumental measurement error, for the purpose of this manuscript we use the word "error-free" to refer to $PM_{2.5}$ concentrations measured at monitors, since these are the best available source for ground-level $PM_{2.5}$ concentrations. Using this internal validation study, we apply our proposed RC-GPS in Section 4. Finally, we conclude with a discussion in Section 5.

2. Methods.

2.1. *General notation and overview.* Let Y denote the observed outcome, X denote the corresponding true continuous exposure, X_c denote the true categorical exposure, which is obtained from X based on pre-specified cut-offs, selected according to scientific interest, W denote the error-prone continuous exposure, W_c denote the error-prone categorical exposure, which is obtained using the same pre-specified cut-offs, $\mathbf{D}$ denote error-free covariates associated with the measurement error, and $\mathbf{C}$ denote error-free confounders associated with the true exposure and outcome. There is no restriction on whether $\mathbf{D}$ and $\mathbf{C}$ include the same covariates or not. For the main study, only $(Y, W, \mathbf{D}, \mathbf{C})$ are observed. In addition, suppose a validation study for which $(X, W, \mathbf{D})$ are observed. Note the validation study does not have to be internal.

Our interest is in estimating the causal effect of a categorical exposure on the outcome in observational studies. The target estimand is the average treatment effect (ATE). Following the potential outcomes framework (Rubin (1974)), we assume no-interference (Cox (1958)), which is sometimes referred to as the stable-unit-treatment-value assumption (SUTVA) (Rubin (1990)). Under this assumption,

we assume that the potential outcome for a given observation is not affected by the exposure of any other unit, and that each exposure defines a unique outcome for each observation.

Furthermore, under overlap and weak unconfoundedness assumptions (discussed in detail later), GPS can be used to estimate the ATE with observed categorical exposures adjusting for confounding. In the main study, the exposure is mismeasured; only W (along with W_c) are observed instead of X_c. Estimating the ATE based on W_c instead of X_c may result in biased estimates of the ATE. Our goal is to adjust for the measurement error in the exposure and obtain unbiased estimates of the ATE. We accomplish this by introducing a regression calibration (RC)-based adjustment for mismeasured exposures combined with GPS to adjust for confounding (RC-GPS). This approach relies on a main study/validation study design.

2.2. *Regression calibration.* In this section, we propose a regression calibration approach to adjust for measurement error in a continuous exposure. The adjustment relies on two common measurement error assumptions: (1) Transportability: we assume that the relationship between X and W, D would be the same in the validation study where X is observed and in the main study in which it is not. (2) Nondifferential measurement error: $Y \perp\!\!\!\perp W | X, \mathbf{D}$. This assumption is equivalent to the surrogacy assumption and it means the conditional distribution of outcome Y given $(W, X, \mathbf{D})$ depends only on $(X, \mathbf{D})$.

The relationship between true exposures X and error-prone exposures W, conditional on other covariates $\mathbf{D}$, is modeled using a regression model specified by mean and variance;

(2.1)
$$E(X|W, \mathbf{D}) = m_X(W, \mathbf{D}, \boldsymbol{\gamma}),$$
$$\text{Var}(X|W, \mathbf{D}) = V(W, \mathbf{D}, \boldsymbol{\gamma}) \Sigma_{X|W, \mathbf{D}} V^T(W, \mathbf{D}, \boldsymbol{\gamma}).$$

Under transportability, we assume that the coefficients $\hat{\boldsymbol{\gamma}}$ which are estimated in the validation study are transportable to the main study. Thus, unobserved X in the main study can be estimated using $m_X(W, \mathbf{D}, \hat{\boldsymbol{\gamma}})$. A well-studied case is a linear regression model specified by:

(2.2)
$$E(X|W, \mathbf{D}) = \gamma_0 + \gamma_1 W + \boldsymbol{\gamma_2}^T \mathbf{D},$$
$$\text{Var}(X|W, \mathbf{D}) = \Sigma_{X|W, \mathbf{D}}.$$

Under transportability, unobserved X in the main study can be estimated using equation 2.2;

$$\hat{X} = \hat{\gamma_0} + \hat{\gamma_1} W + \hat{\boldsymbol{\gamma_2}}^T \mathbf{D}.$$

When $\text{tr}(\Sigma_{X|W, \mathbf{D}}) = 0$, this reduces to the standard regression calibration model in which we only need to estimate $E(X|W, \mathbf{D})$ and $\hat{X}$ is an unbiased estimator of X

(Carroll et al. (2006)). Carroll et al. (2006) proved by Taylor series expansions that $\hat{X}$ is approximately unbiased when the model has good fit, that is, $\text{tr}(\Sigma_{X|W,\mathbf{D}})$ is small.

2.3. *Generalized propensity scores estimation.* In this section, we discuss the generalization of the propensity score, introduced in the causal literature by Rosenbaum and Rubin (1983) for a binary treatment, to the setting of categorical exposures. We follow the generalization proposed by Imbens (2000). Under the assumption that we know the true exposure, denote $X_c \in \mathbb{X}_c = \{1, 2, \ldots, n\}$ the true categorical exposure having n categories. Let $p(x|\mathbf{c}) = \Pr(X_c = x|\mathbf{C} = \mathbf{c})$ for pre-specified categories $x = 1, 2, 3, \ldots, n$. We define the GPS as the conditional probability of receiving each category of the exposure given other pre-exposed covariates $\mathbf{c}$:

$$(2.3) \qquad \text{GPS}(\mathbf{c}) = \big(p(1|\mathbf{c}), p(2|\mathbf{c}), \ldots, p(n|\mathbf{c})\big).$$

The individual $p(x|\mathbf{c})$ is called the xth element of GPS($\mathbf{c}$).

To model GPS($\mathbf{c}$) = $(p(1|\mathbf{c}), p(2|\mathbf{c}), \ldots, p(n|\mathbf{c}))$, we consider a generalized linear model (GLM) relating X_c to $\mathbf{C}$, that is $p(x|\mathbf{c}) = \Pr(X_c = x|\mathbf{C} = \mathbf{c}, \boldsymbol{\eta}) = g^{-1}(\eta_{0x} + \boldsymbol{\eta_{1x}}^T \mathbf{c})$, where g is known. One common g is the multinomial logistic regression model.

$$\ln \frac{\Pr(X_c = x|\mathbf{C} = \mathbf{c}, \boldsymbol{\eta})}{\Pr(X_c = n|\mathbf{C} = \mathbf{c}, \boldsymbol{\eta})} = \eta_{0x} + \boldsymbol{\eta_{1x}}^T \mathbf{c}.$$

2.4. *Generalized propensity scores implementation.* We consider three GPS implementations; subclassification, IPTW and matching, all conditional on the estimated GPS (Imbens (2000), Yang et al. (2016)). Following Yang et al. (2016), let $X_{c,j}$ denote the true categorical exposure for unit j, $X_{c,j} \in \mathbb{X}_c = \{1, 2, \ldots, n\}$, and $Y_j(x)$ denote the potential outcome for an exposure x for observation j. The observed outcome can then be written as $Y_j^{\text{obs}} = Y_j(X_{c,j})$. Define the indicator variables $I_j(x) \in \{0, 1\}$,

$$I_j(x) = \begin{cases} 1 & \text{if } X_{c,j} = x, \\ 0 & \text{otherwise.} \end{cases}$$

In addition to the no-interference assumption described above, we require the following two assumptions for proposing the GPS implementations.

ASSUMPTION 1 (Overlap/positivity). For all values of $\mathbf{c}$, the probability of receiving any category of the exposure is positive:

$$\Pr(X_c = x|\mathbf{C} = \mathbf{c}) > 0 \qquad \text{for all } x, \mathbf{c}.$$

This assumption guarantees that for all possible values of **c**, we will be able to estimate the ATE for each category of the exposure without relying on extrapolation. In many applications, there are regions of the confounder space with low probability values of receiving one of the exposures, which leads to a violation of this assumption. There are methods for improving overlap; specifically, both Harder, Stuart and Anthony (2010) and Yang et al. (2016) suggest dropping units from the analysis with low and high values of the GPS, and conducting analysis on the trimmed sample.

ASSUMPTION 2 (Weak unconfoundedness).   The assignment mechanism is weakly unconfounded if for all $x \in \mathbb{X}_c = \{1, 2, \ldots, n\}$,

$$I_j(x) \perp\!\!\!\perp Y_j(x)|\mathbf{C}_j.$$

There are two things to note about this assumption. First, it can be preserved if we condition on a specific scalar function of $\mathbf{C}_j$, i.e. $p(x|\mathbf{C}_j)$, as shown in Lemma 1 below. This is favorable since it allows for the reduction in the dimension of the conditioning covariates when estimating causal effects. Second, this assumption is sufficient for constructing a form of the ATE, which will be formalized in Lemma 2.

LEMMA 1 (Weak unconfoundedness given GPS).   *Suppose the assignment mechanism is weakly unconfounded. Then for all* $x \in \mathbb{X}_c = \{1, 2, \ldots, n\}$,

$$I_j(x) \perp\!\!\!\perp Y_j(x)|p(x|\mathbf{C}_j).$$

Lemma 1 allows us to estimate the following ATE, described in Lemma 2.

LEMMA 2 (Average treatment effects under weak unconfoundedness).   *Suppose the assignment mechanism is weakly unconfounded. Then for all* $x, x' \in \mathbb{X}_c = \{1, 2, \ldots, n\}$,

$$\text{ATE}(x'; x) = E[Y_j(x') - Y_j(x)]$$
$$= E[E[Y_j^{\text{obs}}|X_{c,j} = x', p(x'|\mathbf{C}_j)]] - E[E[Y_j^{\text{obs}}|X_{c,j} = x, p(x|\mathbf{C}_j)]].$$

Lemma 2 allows us to loosen the constraint in comparison of exposure effects. Instead of conditioning on the full set of $n - 1$ generalized propensity scores $(p(1|\mathbf{C}_j), \ldots, p(n-1|\mathbf{C}_j))$, we can estimate the average effect $E[Y_j(x') - Y_j(x)]$ by constructing an overall average estimate for each exposure category x separately. For a single exposure category x, the corresponding subpopulations are defined by the value of a single score, $p(x|\mathbf{C}_j)$, leading to the equality;

$$E[Y_j(x)] = E[E[Y_j^{\text{obs}}|X_{c,j} = x, p(x|\mathbf{C}_j)]].$$

Even though the comparisons of exposure effects are not constructed by conditioning on the full set of generalized propensity scores, which makes us lose the ability to create subpopulations where we can extrapolate causal effects, our estimated ATE for whole population under the weak unconfoundedness assumption is still valid in causal inference (Imbens (2000)). In contrast, under strong unconfoundedness where $X_{c,j} \perp\!\!\!\perp (Y_j(1), Y_j(2), \ldots, Y_j(n)) | \mathbf{C}_j$ (defined in Rosenbaum and Rubin (1983)), we can estimate ATE for subpopulation, that is, the ATE in population with exposure category 1 and 2 only. Our interest, however, is usually in causal effects for whole population, which can be achieved under the weak unconfoundedness assumption.

2.4.1. *Subclassification.*   We follow the approach proposed by Yang et al. (2016). Consider classifying individuals into K groups based on the xth GPS element, each group containing $N_{k,x}$ observations having similar values of the corresponding estimated GPS elements. The most common way to construct subclasses is to use quantiles of the GPS. The ATE between two exposures, x' and x, that is, ATE$(x'; x)$, can be written as the difference of two expectations $E[Y_j(x')]$ and $E[Y_j(x)]$, which can be estimated separately. Let $q_{x,k}^{p(x|\mathbf{c}_j)}$ be the value of $p(x|\mathbf{C}_j)$ in kth quantile in the sample. The average value of $Y_j(x)$ in subclass k is estimated as;

$$\hat{\mu}_{k,x} = \frac{1}{N_{k,x}} \sum_{j:q_{x,k-1}^{p(x|\mathbf{c})} \leq p(x|\mathbf{C}_j) < q_{x,k}^{p(x|\mathbf{c})}, X_j = x} Y_j^{\text{obs}},$$

where $N_{k,x}$ is the number of units with the xth GPS element falling into the interval $[q_{x,k-1}^{p(x|\mathbf{c})}, q_{x,k}^{p(x|\mathbf{c})})$ and $X_{c,j} = x$. The overall average of $Y(x)$, $E[Y_j(x)]$, is then estimated as,

$$\hat{E}[Y_j(x)] = \hat{E}[E[Y_j^{\text{obs}} | X_{c,j} = x, p(x|\mathbf{C}_j)]]$$

$$= \sum_{k=1}^{K} \frac{N_k}{N} \hat{\mu}_{k,x},$$

where N_k is the number of individuals with the xth GPS element falling into the interval $[q_{x,k-1}^{p(x|\mathbf{c})}, q_{x,k}^{p(x|\mathbf{c})})$, and N is the total sample size. We can estimate $E[Y_j(x')]$ similarly, and consequently, obtain the ATE between any two categories of exposure.

2.4.2. *Inverse probability of treatment weighting (IPTW).*   IPTW involves weighting each individual by the inverse of their GPS. This approach was first introduced by Imbens (2000) and is an analog of using IPTW with propensity scores under a binary exposure. The probability weight assigned to a particular individual is the GPS element corresponding to its true category of treatment. Note that the

use of IPTW can be construed as a further extension of subclassification, with the number of subclasses going to infinity. One can estimate the overall average of $Y_j(x)$ using IPTW as;

$$\hat{E}\big[Y_j(x)\big] = \hat{E}\bigg[\frac{Y_j^{\mathrm{obs}} I_j(x)}{p(x|\mathbf{C}_j)}\bigg].$$

We can estimate $E[Y_j(x')]$ similarly, and consequently, obtain the ATE between any two categories of exposure.

2.4.3. *Matching.* We follow the approach proposed by Yang et al. (2016), which involves matching individuals who receive one category of exposure to individuals who received another category of exposure based on their estimated GPS. There are various ways of matching, for example, matching on the full set of GPS, yet here we match based on a scalar variable to furthest reduce the dimensionality of the matching problem. Following Yang et al. (2016), we define a one-to-one nearest neighbor matching function with replacement,

$$m_{\mathrm{gps}}(x, p) = \arg\min_{j:X_{c,j}=x} \big\| p(x|\mathbf{C}_j) - p \big\|.$$

Using this matching function, we impute $Y_j(x)$ as: $\hat{Y}_j(x) = Y_{m_{\mathrm{gps}}(x, p(x|\mathbf{C}_j))}^{\mathrm{obs}}$ for $j = 1, 2, \ldots, N$ successively, to create a dataset with the sample size N, yet with replicated observations. By resampling with replacement from the original dataset, we can construct a finite sample representing the pseudo subpopulation having exposure $X_{c,j} = x$. The overall average of $Y_j(x)$ can be expressed as;

$$\hat{E}\big[Y_j(x)\big] = \frac{1}{N} \sum_{i=1}^{N} Y_{m_{\mathrm{gps}(x, p(x|\mathbf{C}_i))}}^{\mathrm{obs}}.$$

We can estimate $E[Y_j(x')]$ similarly by creating another dataset with observations receiving exposure $X_{c,j} = x'$ using the matching function defined above, and consequently, obtain the ATE between any two categories of exposure.

2.5. *Outcome analyses.* In the causal framework, one might be interested in a specific statistical quantity, for example, ratio measures. The three GPS implementations are not explicit about the forms of the outcome model, and provide the flexibility to estimate such statistical quantities directly from the estimates of the overall averages for each exposure category. For example, the ATE measured by ratio can be expressed as;

$$\mathrm{ATE}_{\mathrm{ratio}}(x'; x) = \frac{\hat{E}[Y_j(x')]}{\hat{E}[Y_j(x)]}.$$

However, if one is interested in incorporating covariates into the outcome model, to further adjust for confounding, one may want to specify an outcome model. For instance, one may specify the following outcome model; $Y|X_c, \mathbf{C}$ using a GLM,

$E(Y(X_c)) = r^{-1}(\beta_0 + \sum_{x=1}^{n} \beta_{1x} I(X_c = x) + \boldsymbol{\beta_2}\mathbf{C})$, where $I(\cdot)$ is an indicator for the corresponding exposure category. While we conduct all analyses assuming $\beta_2 = 0$ and exclude $\mathbf{C}$ from the outcome model, $\beta_2 \neq 0$ could be included to adjust for residual imbalances not captured by the generalized propensity score implementation or to improve precision of causal estimates (Harder, Stuart and Anthony (2010)).

For IPTW, the outcome model can be easily implemented as a GLM with the corresponding GPS elements specified as weights. For subclassification, the outcome model is essentially implemented on samples selected from subclasses constructed by the corresponding GPS elements, and then weighted by sample size of their corresponding subclass. For matching, the outcome model is essentially implemented on the replicated samples constructed by GPS matching as described above.

For inference, we estimate the standard errors (SEs) of the ATE using bootstrap to jointly account for the variability in the estimation of RC parameters $\boldsymbol{\gamma}$, GPS parameters $\boldsymbol{\eta}$ and outcome model parameters $\boldsymbol{\beta}$. We use standard bootstrap to construct the SEs for GPS subclassfication and IPTW. We use a modified bootstrap method for GPS matching, as standard bootstrap may provide invalid standard errors for matching (Abadie and Imbens (2008)). Certain modifications to the standard bootstrap, like the m-out-of-n bootstrap in Bickel, Götze and van Zwet (2012), were proven to recover the validity.

2.6. *Proposed RC-GPS.* Our proposed RC-GPS adjustment is a two-stage approach.

Stage 1: Measurement error correction

1. Fit a RC model in the validation study. More specifically, fit $E(X|W, \mathbf{D}) = \gamma_0 + \gamma_1 W + \boldsymbol{\gamma_2}^T \mathbf{D}$ to obtain estimated $\boldsymbol{\gamma}$, that is, $\hat{\boldsymbol{\gamma}}$ in the validation study. The form of RC model is not restricted to linear regression, although otherwise one needs further justifications of approximations for the RC model to fully adjust for measurement error (Section 2.2).

2. Under the transportability assumption, estimate $\hat{X} = \hat{\gamma}_0 + \hat{\gamma}_1 W + \hat{\boldsymbol{\gamma}}_2^T \mathbf{D}$ in the main study. The $\hat{X}$ is approximately unbiased if the RC model is correctly specified and has good fit (i.e., $\text{tr}(\Sigma_{X|W,\mathbf{D}}^2)$ is small).

3. Based on pre-defined categories, transform $\hat{X}$ into $\hat{X}_c \in \mathbb{X}_c = \{1, 2, \ldots, n\}$, a categorical variable. The choice of category can be determined to either be policy-relevant or by optimizing overlap through sensitivity analyses.

Stage 2: GPS estimation, implementation, outcome analysis

Stage 2A: Design phase with GPS

4. After obtaining $\hat{X}_c$ in the main study, estimate the GPS model using a GLM relating $\hat{X}_c$ to $\mathbf{C}$ as described in Section 2.3. The estimated GPS is approximately error-free if the RC model is correctly specified and has good fit.

Stage 2B: Analysis phase with GPS

5. Estimate $\hat{E}[Y(x)]$ for each exposure category $x \in \mathbb{X}_c = \{1, 2, \ldots, n\}$ after adjusting for confounding using GPS subclassfication, IPTW or matching methods (Section 2.4).

6. Estimate the ATE as the contrast of $\hat{E}[Y(x)]$ and $\hat{E}[Y(x')]$ between any two exposure categories x, x'.

7. Estimate the SEs of the ATE using bootstrap to jointly account for the variability in the estimation of RC parameters $\boldsymbol{\gamma}$, GPS parameters $\boldsymbol{\eta}$ and outcome model parameters $\boldsymbol{\beta}$.

3. Simulations. We conduct simulation studies to evaluate the performance of the proposed RC-GPS approach under the three types of GPS implementations outlined in Section 2.6 (IPTW, subclassification and matching). We estimate the ATE based on (1) the true exposures in both GPS and outcome models, (2) the error-prone exposures in both GPS and outcome models, (3) our proposed RC-GPS adjustment.

3.1. *Simulation strategies.* We generate a main study/internal validation study setting, in which the validation study is randomly sampled from the main study. The data generation strategy is summarized in Table 1. Briefly, we generate six confounders $(C_1, C_2, \ldots, C_6)$, which include a combination of continuous and categorical variables. We generate three covariates in the measurement error model, (D_1, D_2, D_3), which are continuous. Note that $\mathbf{C}$ and $\mathbf{D}$ could include the same covariates, although this is not required. For simulations we assume that $C_1 = D_1$, but that the remaining covariates are different.

The variables $[W|\mathbf{C}, \boldsymbol{\tau}], [X|W, \mathbf{D}, \boldsymbol{\gamma}], [Y|X, \mathbf{C}, \boldsymbol{\beta}]$ were generated as continuous under linear regression models with parameters specified in Table 1. We begin by generating the error-prone exposures W, and then generate X, which guarantees the correct specification of the RC model. We consider seven settings, where we vary (1) $\boldsymbol{\tau}$ to control the strength of confounding for exposure, (2) γ_1 to control the correlations between X and W, (3) $\Sigma_{X|W, \mathbf{D}}$ to control the goodness of RC model fit, (4) quadratic term in the RC model to control RC model misspecification, (5) β_1 to control the magnitude of treatment effect in outcome model, and (6) $\boldsymbol{\beta}_2$ to control the strength of confounding for outcome. The default setting, discussed in detail in the main text, is highlighted in Table 1. We fix the sample size of the main study as 2000 and the internal validation study as 500. We conduct 1000 replicates of each scenario. The R code for all simulations is available on github https://github.com/wxwx1993/RC-GPS.

3.2. *Simulation results.* To implement the RC-GPS we follow the approach described in Section 2.6. After fitting the RC model and estimating $\hat{X}$ in the main study, we categorize these estimates into three categories based on pre-defined

TABLE 1
Simulation parameters: *data generating mechanism under which simulations were conducted*

Confounding for exposure, $E[W \mid \mathbf{C}, \boldsymbol{\tau}]$	$\boldsymbol{\tau}$
(1) **Moderate confounding**	$(0.8, 0.8, 1.6, 1.2, 2.4, 1.6, 2.4)^T$
(2) Large confounding	$(1.6, 1.6, 3.2, 2.4, 4.8, 3.2, 4.8)^T$
Measurement Error Model, $E[X \mid W, \mathbf{D}, \boldsymbol{\gamma}]$ $\boldsymbol{\gamma}$	
(1) **Strong correlation: 0.85**	$\gamma_1 = 0.8$, $\boldsymbol{\gamma}_2 = (2, 1, 3)^T$
(2) Weak correlation: 0.40	$\gamma_1 = 0.2$, $\boldsymbol{\gamma}_2 = (2, 1, 3)^T$
Measurement Error Model Fit	$\Sigma_{X \mid W, \mathbf{D}}$
(1) **Good of fit**	$\mathrm{diag}(\Sigma_{X \mid W, \mathbf{D}}) = \boldsymbol{I}_N$
(2) Lack of fit	$\mathrm{diag}(\Sigma_{X \mid W, \mathbf{D}}) = 10 \boldsymbol{I}_N$
Measurement Error Model Specification	Model structure
(1) **Linear Model**	$X = \gamma_1 W + \boldsymbol{\gamma}_2 \mathbf{D}$
(2) Quadratic Model ($\gamma_3 = 0.05$)	$X = \gamma_1 W + \boldsymbol{\gamma}_2 \mathbf{D} + \gamma_3 W^2$
Outcome Model, $E[Y \mid X, \mathbf{C}, \boldsymbol{\beta}]$	$\boldsymbol{\beta}$
(1) **Large treatment effect**	$\beta_1 = 1$, $\boldsymbol{\beta}_2 = (3, 2, 1, 4, 2, 1)^T$
(2) Small treatment effect	$\beta_1 = 0.5$, $\boldsymbol{\beta}_2 = (3, 2, 1, 4, 2, 1)^T$
Confounding for outcome	$\boldsymbol{\beta}$
(1) **Moderate confounding**	$\beta_1 = 1$, $\boldsymbol{\beta}_2 = (3, 2, 1, 4, 2, 1)^T$
(2) Large confounding	$\beta_1 = 1$, $\boldsymbol{\beta}_2 = (15, 10, 5, 20, 10, 5)^T$
Sample size	N_m and N_v
(1)	$N_m = 2000$ and $N_v = 500$
Covariate	Distribution
(1) $C_1 - C_3/C_4/C_5/C_6$	$N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 & -1 \\ 1 & 1 & -0.5 \\ -1 & -0.5 & 1 \end{pmatrix}\right] / U\{-2, 2\} / U(-3, 3) / \chi^2(1)$
(2) $D_1/D_2/X_3$	$C_1 / N(0, 4) / U(-5, 5)$
Cut-off points	$\mathbf{k}$
(1)	$k_1 = -5$ and $k_2 = 15$

cut-off points ($k_1 = -5$, $k_2 = 15$), and obtain exposure categories $\hat{X}_c = 1, 2, 3$. We also obtain W_c from W using the same cutoffs. Using $\hat{X}_c$ we then fit the GPS model and estimate the ATE. For subclassification, we classify subjects into 10 subclasses by deciles based on each GPS element. For IPTW, weights were calculated as the inverse of the corresponding GPS elements as described in Section 2.4, and extreme weights are set equal to 10 if the weights are greater than 10, named "truncating" (Harder, Stuart and Anthony (2010)). For matching, we use a form of one-to-one nearest neighbor matching with replacement based on the corresponding GPS elements as described in Section 2.4.

We provide a detailed description of the simulation results from the default setting (highlighted in bold in Table 1). Under this setting, the true ATE is $\boldsymbol{\beta_1} =$
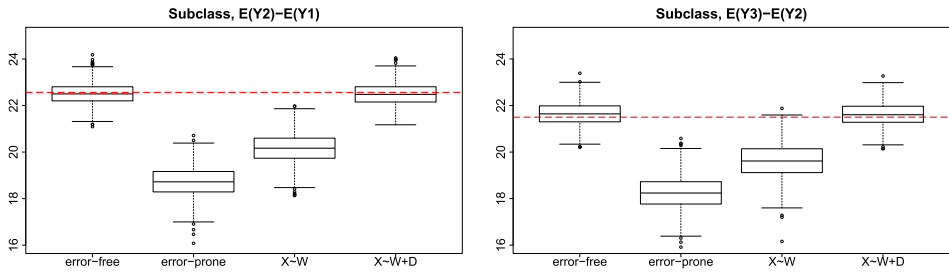
FIG. 1. *Simulation results for default setting described in Table* 1. *Subclassification*: (a) *Error-free*: *GPS approach using error-free exposures* X_c; (b) *Error-prone*: *GPS approach using error-prone exposure* W_c; (c) $X \sim W$: *the proposed two-stage RC-GPS approach, in which exposures are estimated using a misspecified RC model* $\hat{X} = \hat{\gamma}_0 + \hat{\gamma}_1 W$ *which does not include covariates*; (d) $X \sim W + D$: *the proposed two-stage RC-GPS approach, in which exposures are estimated using the correctly specified RC model* $\hat{X} = \hat{\gamma}_0 + \hat{\gamma}_1 W + \hat{\boldsymbol{\gamma}_2}^T \mathbf{D}$ *which includes covariates. The red dashed line represents the true ATE* (*gold standard*).

$(22.56, 21.50)$ (for exposure categories $\hat{X}_c = 2$ vs. 1 and $\hat{X}_c = 3$ vs. 2, respectively) which is estimated by fitting the linear model $Y = \beta_0 + \sum_{i=2}^{3} \beta_{1i} I(X_c = i) + \boldsymbol{\beta_2} \mathbf{C}$ for a large simulated dataset with sample size $N = 10^6$.

The ATE for subclassification is shown in Figure 1. The left plot represents the ATE of exposure $\hat{X}_c = 2$ vs. $\hat{X}_c = 1$, and the right plot represents the ATE of exposure $\hat{X}_c = 3$ vs. $\hat{X}_c = 2$. The ATE is estimated based on four different approaches from left to right: (a) based on GPS approach using error-free exposure X_c categorized from X; (b) based on GPS approach using error-prone exposure W_c categorized from W; (c) based on the proposed two-stage RC-GPS approach, in which $\hat{X}$ is estimated using a misspecified RC model $\hat{X} = \hat{\gamma}_0 + \hat{\gamma}_1 W$ which does not include covariates; (d) based on proposed two-stage RC-GPS approach, in which $\hat{X}_c$ is estimated using the correctly specified RC model $\hat{X} = \hat{\gamma}_0 + \hat{\gamma}_1 W + \hat{\boldsymbol{\gamma}_2}^T \mathbf{D}$ which includes covariates. The true ATE is denoted by the red dashed line.

GPS implementation with subclassification using the error-free exposures results in a very small bias compared to the true ATE. Yet even in this setting where the error-prone and error-free exposures are highly correlated, GPS implementations using error-prone exposures result in significant bias, which illustrates the necessity of adjusting for the measurement error. RC-GPS using a correctly specified RC model performs really well, significantly reducing the bias of the estimated ATE compared to using the error-prone exposure. The bias was reduced from -17.07% to -0.36% and from -15.13% to 0.55% (exposure categories $\hat{X}_c = 2$ vs. 1 and $\hat{X}_c = 3$ vs. 2, respectively). RC-GPS using a misspecified RC model still reduces, although does not completely eliminate, the bias. The bias under this setting was reduced from -17.07% to -10.58% and from -15.13% to -8.77% (exposure categories $\hat{X}_c = 2$ vs. 1 and $\hat{X}_c = 3$ vs. 2, respectively). Additional results (Figure A1 in Supplement (Wu et al. (2019))), show that IPTW and matching perform similarly compared to subclassification.
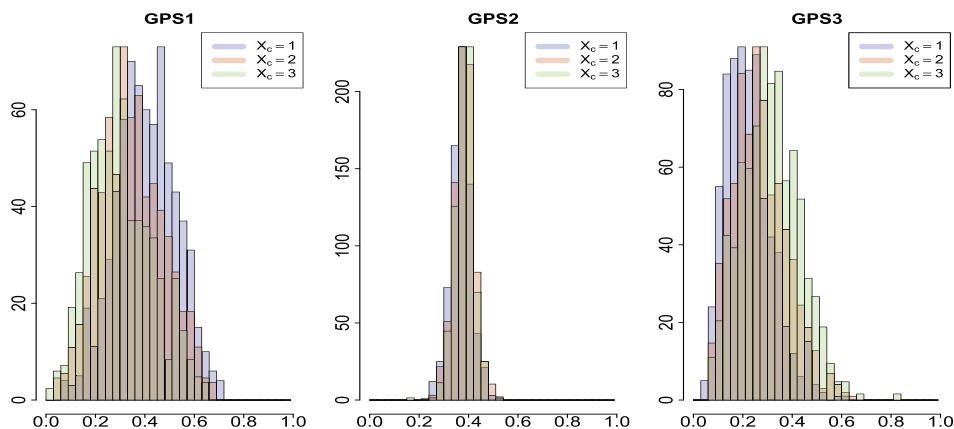
FIG. 2. *Overlap. Each panel represents histograms of each corrected GPS element, colored according to different subpopulations. For assessing overlap, we can see the majority of samples have GPS elements away from zero or one, and the ranges of each GPS element overlap across samples with different exposure providing the evidence that the overlap assumption likely holds. Results corresponding to the default setting are described in Table* 1.

3.3. *Overlap and balance.* We evaluate the overlap assumption by inspecting the distributions of each estimated GPS element for all subjects as shown in Figure 2. In this simulation study, we see that the overlap assumption likely holds in general, since the majority of samples have GPS elements away from zero or one. The histogram also shows the ranges of each GPS element overlap across samples with different exposure, which is referred to complete overlap in practice (Vaughn (2008)). Therefore, we do not need to trim the data. In additional simulations in which the overlap assumption is (close to being) violated, we trimmed the data to improve overlap. For IPTW, we also minimize the influence of outlier weights by rounding down extreme weights, also referred to as truncation. We see that the estimates based on the three GPS implementations become less consistent. The performance of IPTW is notably different from the other two approaches. This difference is due to the altering of the target estimand which is caused when truncating the weights for a potentially large portion of observations. We also find the untruncated IPTW estimates provide overall less biased results yet may have large variances (outliers) compared to truncated IPTW estimates (Figures A11–13 in Supplement (Wu et al. (2019))).

Under weak unconfoundness, the single GPS element $p(x|\mathbf{c_j})$ is only required to achieve balance between subpopulations with $X_c = x$ and subpopulations with $X_c \neq x$. One can assess balance by estimating the absolute standardized biases of each confounder before and after GPS implementation using techniques similar to those described in Harder, Stuart and Anthony (2010). The key is that each GPS element $p(x|\mathbf{c_j})$ is treated as a binary exposure propensity score, and balance is evaluated across all confounders between subpopulations with $X_c = x$ and
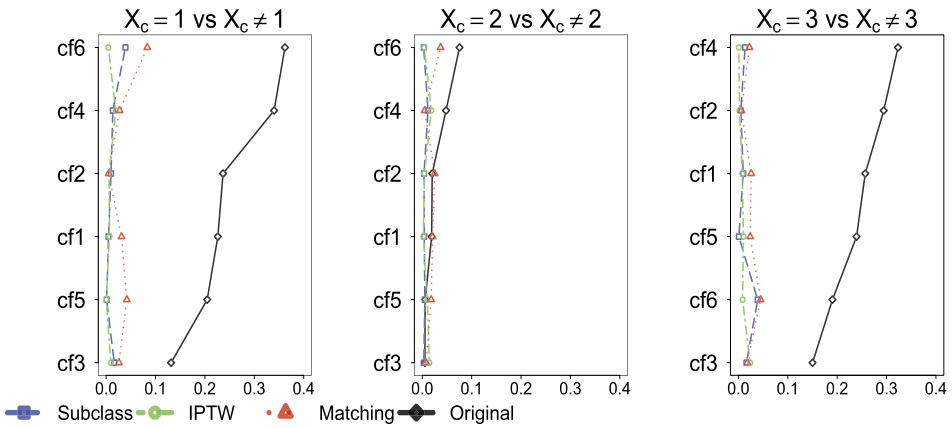
FIG. 3. *Absolute standardized bias (ASB). Each panel represents the absolute standardized biases for each of the six confounders (cf), between subpopulations with $X_c = x$ and subpopulations with $X_c \neq x$ in the original data (solid line) and after GPS implementations (dashed line). All three GPS implementations perform similarly and all improve confounder balance substantially. Results corresponding to the default setting are described in Table* 1.

subpopulations with $X_c \neq x$. For each of the six confounders, we estimate the absolute standardized bias (ASB). The ASB for each covariate is calculated by dividing the difference in means of the covariate between the treated group and the comparison group by the standard deviation (Harder, Stuart and Anthony (2010)). We see that balance improves substantially across all six confounders for all three implementation approaches (Figure 3).

3.4. *Sensitivity analysis.* We assess the sensitivity of the proposed approach to the transportability assumption by evaluating how well the approach performs under settings in which the true γ is misspecified. We sample $\widehat{\gamma_{1,a}}$ from a normal distribution with mean $\hat{\gamma}_1$ and augmented standard deviation estimated by adding absolute values 0.1/0.2/0.3/0.5 to the original standard deviation (0.0023) of $\hat{\gamma}_1$. $\hat{X}$ is then estimated by $\hat{X} = \hat{\gamma}_0 + \widehat{\gamma_{1,a}} W + \hat{\boldsymbol{\gamma}_2}^T \mathbf{D}$. By adding this misspecification of the RC model, we artificially violate the transportability assumption and show how this violation could affect the ATE estimates. The results in Figure 4 show that for subclassification even when the standard deviation is around 100 times higher than the original standard deviation, the estimated ATE using GPS is still robust. Not surprisingly, the variances of the estimated ATE increase under this extreme setting. The results from using GPS with IPTW and matching are also robust to the violation of transportability assumption (see Figure A2 in Supplement (Wu et al. (2019))).

3.5. *Additional simulations.* We conducted additional simulations (as described in Table 1) with varying degrees of exposure error and confounding bias.
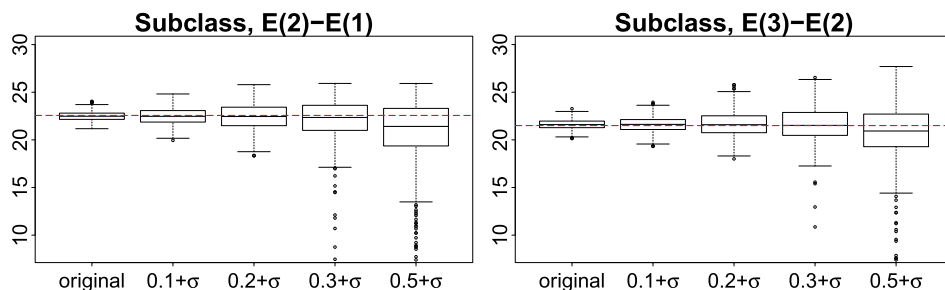
FIG. 4. *Sensitivity analysis of the ATE estimates based on GPS approach using subclassification. Original represents the estimated ATE when the transportability assumption holds. Various violations of the transportability assumption are conducted, by sampling γ in the RC model from a distribution with mean $\hat{\gamma}$ and augmented standard deviation. The red dashed line represents the true ATE. Results correspond to the default setting described in Table* 1.

Details on the additional simulations can be found in the Supplement (Wu et al. (2019)). Briefly, we show that in a variety of settings our proposed RC-GPS approach can significantly reduce the bias of the estimated ATE. More specifically, when the correlation between true exposure X and error-prone exposure W is low the proposed approach significantly eliminates the bias if the RC model is correctly specified and has good fit. Yet the proposed approach is more sensitive to the correct specification of the RC model in these settings (see Figure A3 in Supplement (Wu et al. (2019))). When the RC model does not fit well the proposed method improves the bias in the ATE, but does not completely eliminate it (see Figure A4 in Supplement (Wu et al. (2019))). When the true treatment effect is small the proposed approach eliminates the bias if the RC model is correctly specified and has good fit (see Figure A5 in Supplement (Wu et al. (2019))). When confounding is large the proposed approach significantly eliminates the bias if the RC model is correctly specified and has good fit, yet the variance of the estimated ATE increases (see Figures A6 and A7 in Supplement (Wu et al. (2019))). Lastly, when the RC model is nonlinear with respect to error-prone exposure W the proposed approach significantly eliminates the bias if the RC model is correctly specified and has good fit. Yet if we only use a linear model as our RC model (when the data generating mechanism was from a nonlinear model), we do not see bias reductions using the proposed approach (see Figure A8 in Supplement (Wu et al. (2019))). Overall, through the simulation study, we show the RC-GPS approach works remarkably well with varying degrees of exposure error and confounding bias.

4. Data application. We apply the proposed RC-GPS method to estimate the effect of long-term $PM_{2.5}$ exposure on health outcomes. While $PM_{2.5}$ concentrations are continuous, our interest is in comparing the effects of exposure in three categories based on pre-specified $PM_{2.5}$ cut-offs. The current long-term $PM_{2.5}$

standard is annual mean of 12.0 $\mu g/m^3$, refer to National Ambient Air Quality Standards (NAAQS) Table (USEPA (2012)) There is a wide literature studying the effect of $PM_{2.5}$ exposures at these higher levels (Beelen et al. (2014), Dockery et al. (1993), Kioumourtzoglou et al. (2016)), yet limited literature is available on the effects of exposure at lower levels (Shi et al. (2016), Villeneuve et al. (2015)).

Our interest is in estimating the exposure effects in the lower ranges, the results of which can help inform future policy regulations. Specifically, we consider two cut-offs; annual mean $PM_{2.5}$ levels of 8 and 10 $\mu g/m^3$, resulting in three exposure categories. Our main study population is Medicare participants across New England (VT, NH, CT, MA, RI and ME) from 2000 to 2012, and all-cause mortality is the outcome of interest. This study population includes a total of 3.3 million individuals with 24.5 million person-years of follow up, who reside in 2202 zip codes.

$PM_{2.5}$ exposures are determined at each 1 km $\times$ 1 km grid cell using a spatio-temporal prediction model which uses multiple different sources as input (Di et al. (2016)). Although the prediction model performs well (Di, Koutrakis and Schwartz (2016)), there is still error associated with these predictions. For a subset of grid cells we have monitor stations that measure the actual observed $PM_{2.5}$ concentrations. We assume $PM_{2.5}$ concentrations monitored inside a grid cell are error-free exposures of the average concentrations in that grid cell, which is a reasonable assumption (Burton, Suh and Koutrakis (1996), Sarnat et al. (2010), Wilson and Suh (1997)). For New England we have 2202 zip codes covered by 217,660 1 km $\times$ 1 km grids (main study). For a subset of these grids ($m = 83$) within 75 zip codes, we have actual $PM_{2.5}$ measured from monitoring stations (internal validation study). Figure 5 shows the locations of all 119 monitor stations in New England, 83 of which have actual $PM_{2.5}$ measures. Medicare data is 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 through area-weighted averages. The distributions of annual mean $PM_{2.5}$ exposures from the spatio-temporal prediction model in the main and the validation studies are compared in Figure 6, showing that the monitors are not randomized across areas, that is, they are more likely to be located at areas with higher $PM_{2.5}$ concentrations.

RC model. The RC stage of our two-stage RC-GPS approach is implemented at the grid level. We have 217,660 grids in the main study and 83 grids in the internal validation study. After fitting the RC model, we obtained estimates of the true $PM_{2.5}$ exposures at each grid cell in New England. To improve the fit of the RC model we included 14 meteorological variables as predictors, many of which were significant, with total cloud coverage ($p < 0.001$) and total precipitation ($p = 0.008$) as the most significant ones. The details of model fit are presented in the Supplement (Wu et al. (2019)). Subsequently, to obtain annual average $PM_{2.5}$ at the zip code, we aggregated the grid-level $PM_{2.5}$ exposures using area-weighted averages. After the aggregation, we categorized the exposures into three categories,
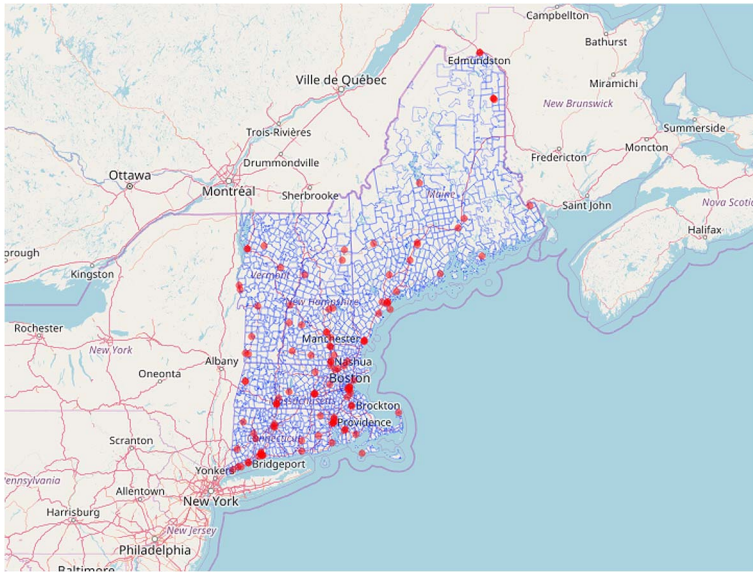
FIG. 5. *Locations of monitor stations in New England (in red). Zip code areas are drawn in blue.*

corresponding to exposure levels $PM_{2.5} \leq 8 \ \mu g/m^3$, $8 < PM_{2.5} \leq 10 \ \mu g/m^3$ and $PM_{2.5} > 10 \ \mu g/m^3$. For each calendar year that participants were at risk, their exposure was the annual average $PM_{2.5}$ for that year, based on their zip code of residence.

GPS model. For the GPS model we include 16 area-level covariates as confounders. The GPS model is fitted using multinomial logistic regression with the
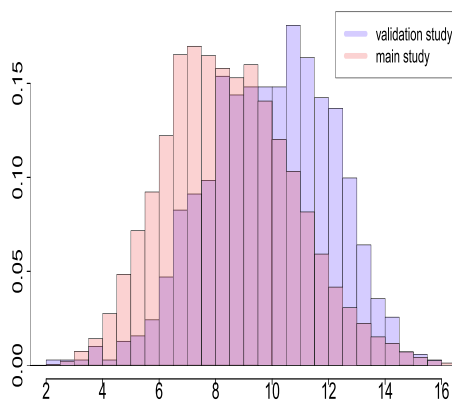

FIG. 6. *The distribution of annual mean predicted* $PM_{2.5}$ *exposures in the main and the validation study across* 13 *years* (2000–2012).

16 confounders. The details of the model fit are presented in the Supplement (Wu et al. (2019)).

Outcome analysis. Following the GPS implementations, we fit the outcome model using a stratified log-linear model with a person-time offset. We use 4 individual level covariates as stratification variables. We do not include confounders in the outcome model and assume the GPS implementations fully adjust for all potential confounding. The incidence rate ratio (IRR) is estimated from the outcome model, and 95% confidence intervals (CIs) were obtained by bootstrap with 100 replicates. We constructed the CIs using standard errors (SEs) estimated by bootstrap under the normality assumption, since estimation of a SE requires fewer bootstrapped replicates (25–200) than the direct estimation of the CI (1000–2000) (Efron and Tibshirani (1994)). We conducted 100 replicates in the data application.

4.1. *Data analysis results.* For each of the GPS implementations, we compare the estimated IRR using (1) a GPS approach using error-prone $PM_{2.5}$ exposures only, and (2) the proposed RC-GPS approach. To improve overlap, we trim the data to include only observations with GPS falling into the overlapping ranges of each GPS element among the different exposure subpopulations, which removes 1.7% of the original data. For IPTW truncating, we further set weights equal to 10 if the weights are greater than 10 (Harder, Stuart and Anthony (2010)), which truncates the weights of 3.1% of the observations. We conduct the outcome analysis based on the trimmed dataset.

We see in Table 2 that the IRR estimates from the RC-GPS approach are consistent across all three implementations. The RC-GPS approach yields more pronounced point estimates compared to the error-prone implementation. The 95% CIs are overlapping across all three implementations. For example, using matching we see a IRR of 1.028 under RC-GPS, for exposure category 1 vs. category 2, meaning a moderate exposure level of annual average $PM_{2.5}$ ($8 < PM_{2.5} \leq 10$ $\mu g/m^3$) causes a 2.8% increase in all-cause mortality compared to low exposure level of annual average $PM_{2.5}$ ($PM_{2.5} \leq 8$ $\mu g/m^3$). Using error-prone exposures we see less consistent results across the three different GPS implementations. The difference in results across the three GPS implementations indicates that these three approaches have different levels of sensitivity to measurement error. It is worth noting that truncating weights in IPTW will affect the estimate of ATE itself (Harder, Stuart and Anthony (2010)), thus possibly causes the differences in results between IPTW and the other two approaches. The untruncated IPTW guarantees asymptotically unbiased estimates of the ATE, yet is sensitive to extreme weights, and performs poorly both in terms bias and variance in finite samples. Therefore, in the data application we recommend careful thought when truncating weights for IPTW.

We assess overlap by evaluating the distributions of the GPS elements for each exposure category as discussed in Section 3.3. Comparing the figures before and

TABLE 2

Data application results: ATE of long-term $PM_{2.5}$ *exposure on mortality measured by incidence rate ratios (IRRs). Error-prone implements GPS approaches to adjust confounding based on error-prone exposures. RC-GPS is based on the proposed approach adjusting for measurement error by RC model and adjusting confounding using GPS approaches based on corrected exposures. All 95% confidence intervals were obtained by bootstrap*

|	ATE [95% CI]			
	Subclassification	**IPTW**	**Matching**	**IPTW (untruncated)**
Results for exposure levels $PM_{2.5} \leq 8$ $\mu g/m^3$ vs. $8 < PM_{2.5} \leq 10$ $\mu g/m^3$				
GPS, Error-prone	1.013 [0.999, 1.029]	1.031 [1.021, 1.042]	1.020 [1.004, 1.036]	0.963 [0.894, 1.038]
RC-GPS	1.025 [1.006, 1.045]	1.022 [1.007, 1.038]	1.028 [1.012, 1.045]	1.032 [0.978, 1.088]
Results for exposure levels $PM_{2.5} \leq 8$ $\mu g/m^3$ vs. $PM_{2.5} > 10$ $\mu g/m^3$				
GPS, Error-prone	1.015 [0.993, 1.037]	1.050 [1.032, 1.068]	1.018 [0.996, 1.040]	0.981 [0.892, 1.079]
RC-GPS	1.035 [0.999, 1.072]	1.030 [1.005, 1.056]	1.035 [1.015, 1.055]	1.043 [0.973, 1.118]
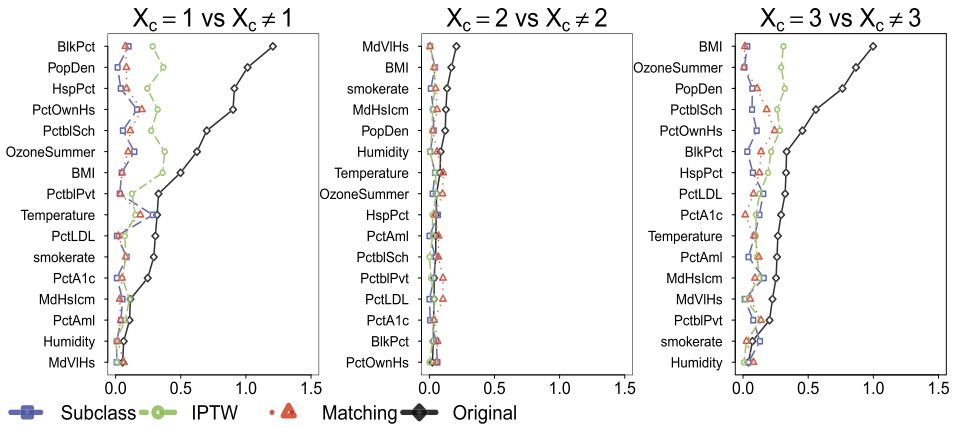
FIG. 7.   *Absolute standardized bias (ASB). Each panel represents the absolute standardized biases for each confounders, between subpopulation with $X_c = x$ and subpopulation with $X_c \neq x$ in original data (solid line) and after GPS implementations (dashed line). All three GPS implementations improve the covariates balances for most of confounders.*

after trimming (see Supplement (Wu et al. (2019))), we notice the overlap assumption does not hold using the original data, but improves after trimming. This highlights the necessity of trimming in order to improve overlap. We assess the balance by calculating the ASB for each confounder for the different exposure categories (e.g., $X_c = 1$ vs. $X_c \neq 1$) as discussed in Section 3.3. It is also important to note, that based on the calculations and figures the GPS implementations largely improve covariate balance for most of confounders (see Figure 7 and Table A10 in Supplement (Wu et al. (2019))).

5. Discussion.   We developed an innovative two-stage approach, RC-GPS, to estimate the average causal effect on a categorical scale in the setting of GPS analysis while correcting for measurement error in continuous exposures. Our simulation study shows that the proposed method has the potential to fully adjust for both the mismeasured exposure as well as confounding bias. We have also conducted sensitivity analyses and showed that the approach is robust under modest levels of model misspecification and assumption violations.

The assumptions for the first component of our proposed approach, the RC model, are (1) transportability and (2) nondifferential measurement error (surrogacy). In our setting, we require the transportability of $E(X|W, \mathbf{D})$. Although this is not verifiable, it can be evaluated by sensitivity analysis, which we included as part of our simulation study in Section 3. We see that in simulation scenarios considered, results are robust to the violation of this assumption. For the application, one should give careful thought about how likely this assumption will hold, and a sensitivity analysis to assess how the ATE varies for the violation of transportability assumption is recommended. The nondifferential measurement error (surro-

gacy) assumption is believed to be held in many similar settings in air pollution studies (Dominici, Zeger and Samet (2000), Hart et al. (2015), Kioumourtzoglou et al. (2014)). We believe it holds, since in the context of air pollution applications, it is reasonable to assume that given the true exposures and confounders, the error-prone exposures do not provide any additional information on the health outcomes (i.e., mortality). In our study, we only consider the RC model as a linear regression model, though the form of RC model is not restricted to linear regression. However, since the RC model is only an approximation, fully adjusting for measurement error relies on assumptions, for example, measurement error is "small" (Carroll and Stefanski (1990)), or outcome models without severe curvature (Carroll et al. (2006)), if other forms of the RC model are used.

The regression of X on $(W, \mathbf{D})$, is an art (Carroll et al. (2006)), since true model of $X|W, \mathbf{D}$ can never be known. Therefore, assessing the robustness of the RC to model misspecification is very important. In summary, we considered three types of model misspecification for the RC model (2.2) in simulations: (1) vary γ in a correctly specified RC model, (2) exclude covariates D associated with the measurement error in the RC model, (3) vary the true model structure of the RC model by introducing a nonlinear (quadric) relationship between true exposure X and error prone W. The first type of model misspecification introduces additional variability in estimating the coefficients γ, although it does not violate the true structure of the RC model (2.2). Under this type of misspecification, we found that the RC-GPS maintains the capability to eliminate the bias, yet the variances of the estimated ATE increase. The second type actually violates the specification of the RC model (2.2) since omission of the covariates $\mathbf{D}$ changes the mean function. We found that under this type of model misspecification, the RC-GPS approach still reduces the bias, although it does not completely eliminate it. The third type (shown in Figure A8 in Supplement (Wu et al. (2019))) is a severe misspecification, since we introduce a nonlinear relationship between X and W, which we ignore when we fit the RC model. We present this type of extreme model misspecification to show that under such a severe case, bias reduction is not guaranteed. Correct specification of the RC model, however, would still eliminate the bias. It is highly recommended, therefore, that goodness of fit is assessed and sensitivity analyses performed to best characterize the functional form of the RC model in real-life applications.

For the GPS implementation, there are three main assumptions (1) no-interference, (2) overlap and (3) weak unconfoundedness. The no-interference assumption is a fundamental assumption in the potential outcome framework. In the air pollution context it could be violated as exposure in the current period could affect mortality in subsequent periods (Baccini et al. (2017)). Baccini et al. (2017) argue that by enlarging the time window of exposure averages the no-interference assumption is more likely to hold. In the data application we consider long-term annual mean exposures (rather than short-term daily exposures), which will likely increase the validity of the assumption.

As we saw in the data application, there are settings in which the overlap assumption does not hold. A common approach, and also the approach we used in data application, to improve overlap is trimming the sample by disregarding subjects with low and high values of GPS elements (Crump et al. (2009)). Yet the limitation is that by doing so we alter the target population, and thus the target estimand. The estimated ATE based on the trimmed sample, can deviate from the quantity of interest, that is, the ATE for the whole population, thus we need to carefully think about the generalizability of our results. In our data application, 1.7% of the observations were trimmed. We compare population characteristics in the entire population and trimmed population, and there is no evidence that those two populations are significantly different (see Table A8 in the Supplement (Wu et al. (2019))).

The unconfoundedness assumption is not verifiable, since data is always uninformative about the distribution of counter-factual outcome for unreceived exposures, yet this is a common assumption in propensity score analysis. In our setting, we only assume weak unconfoundedness, which only requires the potential outcome for each category of exposure and the exposure to be assigned at the corresponding category, are pairwise independent conditional on all potential confounders (Imbens (2000)). It is weaker than strong unconfoundedness defined in Rosenbaum and Rubin (1983), which requires the joint distribution of potential outcomes to be independent with the assignment mechanism for all exposures conditional on all potential confounders. The limitation of weak unconfoundedness is that we are not able to estimate ATE for subpopulations, that is, the ATE in populations with exposure category 1 and 2 only. However, the interest is usually in estimating the ATE for the whole population, which can be estimated under weak unconfoundedness.

For inference, the main strategy is to use bootstrapping to obtain the CIs, and the validity of inference is guaranteed by the validity of the bootstrap procedures. The RC stage in most scenarios does not introduce a substantial amount of additional variability in effect estimates, except when both the RC model lacks fit and the validation size is small (see Tables A2–A3 in Supplement (Wu et al. (2019))). In the data application, the results show (slightly) wider CIs for RC-GPS compared to the naive GPS estimates which do not consider the RC correction (Table 2). The reason that we obtain wider CIs for RC-GPS compared to GPS without correction is that: (1) the validation size is relatively small, (2) the propagation of the uncertainty in the RC stage, (3) in our application, Medicare data is available at the zip code level, but $PM_{2.5}$ exposures are estimated at the grid level. To obtain annual average $PM_{2.5}$ at each zip code, we aggregate these gridded concentrations through area-weighted averages. The aggregation procedure itself could potentially introduce a lot variability as well. Explicitly, the aggregation procedure may amplify the uncertainty during the estimation of RC models. The GPS estimation stage, in general, does not increase the variability in effect estimates compared to those using the true GPS, since Abadie and Imbens (2016), Lunceford and Davidian

(2004) proved that for all three types of PS implementations using the estimated PS is more efficient than using the true PS in large samples, and the arguments in Imai and van Dyk (2004) confirmed the conclusion can be extended to GPS settings. We found similar finite sample properties in simulations (see Tables A4–A5 in Supplement (Wu et al. (2019))).

This data application illustrates the ability of the proposed RC-GPS approach to estimate corrected causal effects between long-term $PM_{2.5}$ on a categorical scale and all cause mortality. There are a few potential limitations in our analysis of the data application. The first is that the validation data is not a random sample of main study, since monitor locations are not randomized across areas, that is, more monitors are within urban areas, which could impact the transportability assumption. However, by conditioning on additional covariates in the RC model, such as weather variables which explain geographical heterogeneity between grid cells, the transportability assumption is more likely to hold. The second is that the RC model is not guaranteed to be correctly specified or have a good fit, due to both lack of potential predictive covariates and modeling assumptions. The third is that the estimated GPS can be biased since it relies on correct specification of the GPS model, and is not robust to unmeasured confounding. The correctly specified GPS maintains a balancing property as described in Section 3.3. In the data application, we assessed balance and there was some evidence of imbalance. Even with some evidence of imbalance, the advantage of GPS approaches is that they are more robust to outcome model misspecification compared to fitting outcome models with confounders as covariates. The fourth is that overlap may still be limited even after we trim the data. There is always a trade-off between guaranteeing overlap and trimming out samples excessively, thus modifying the population. The fifth is the cut-off selection. In the data application, cut-offs were selected from a policy perspective. The first $PM_{2.5}$ cut-off was selected at 10 $\mu g/m^3$, which is the current Air Quality Guideline proposed by the World Health Organization (WHO) for annual $PM_{2.5}$ concentrations (WHO (2018)). Currently, in the US the NAAQS is 12 $\mu g/m^3$ (USEPA (2012)). Understanding, therefore, the effect of $PM_{2.5}$ exposure on mortality in the US population at lower levels, like the WHO guideline or even lower, is of great interest, as it can inform regulatory action. This is also the motivation behind including an additional cut-off at 8 $\mu g/m^3$ annual mean. From a statistical perspective, potential cut-off choices could be driven by modeling assumptions. Specifically, cut-offs can be selected to ensure that the overlap assumption holds for valid causal inference. Based on the distribution of annual mean $PM_{2.5}$ exposures, the chosen 8 $\mu g/m^3$ and 10 $\mu g/m^3$ cut-offs divide the units into three categories approximately evenly, which arguably is more likely to ensure overlap. We have also conducted some sensitivity analysis to evaluate the overlap assumption (see Table A9 in the Supplement (Wu et al. (2019))); the current cut-off points (8 $\mu g/m^3$ and 10 $\mu g/m^3$) provided the best overlap.

Although air pollution has motivated our application, the proposed RC-GPS approach is not limited to one specific area. One other potential application of this

approach could be in the setting of clinical studies of biomarkers. Commonly, in clinical trials, we are interested in the dose response of categorical levels of drug doses, for example, Vitamin D supplements, yet in order to evaluate this treatment in human subjects, we measure the biomarkers, for example, blood levels of Vitamin D. In this setting we might have an error-prone continuous treatment for each patient based on blood work from routine medical examinations, and an internal subset of samples for which we know the true treatment based on blood work from a more robust central laboratory. The accurate measures of blood samples through a standard central laboratory are costly and infeasible for every patient, and therefore we obtain this gold-standard measurement only for a subset of patients (internal validation study). A detailed example can be found in Gail et al. (2016). In such study designs, one can fit a RC model to estimate true treatments for each patient. After that, one can use GPS based on the estimated true treatments in the main study, in order to obtain causal effects for various vitamin D doses, to determine the most effective dose of Vitamin D supplements.

The RC-GPS approach introduced in this paper is the first approach which allows for the correction of exposure error in both design and analysis phases using GPS, and assesses covariate balance through standardized bias in the context of GPS for categorical exposures. It can be further generalized to estimate causal effects on a continuous scale rather than categorical to answer different scientific questions. Simulations have shown the proposed approach is robust and we are optimistic about the adaption of the approach to various research areas.

SUPPLEMENTARY MATERIAL

Supplement to "Causal inference in the context of an error prone exposure: Air pollution and mortality" (DOI: 10.1214/18-AOAS1206SUPP; .pdf). We provided supplementary figures, tables and text that show results from additional simulation scenarios and more details of the data application. R code for simulations is available at https://github.com/wxwx1993/RC-GPS.

REFERENCES

ABADIE, A. and IMBENS, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica* **76** 1537–1557. MR2468559

ABADIE, A. and IMBENS, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84** 781–807. MR3481379

ALEXEEFF, S. E., CARROLL, R. J. and COULL, B. (2016). Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures. *Biostatistics* **17** 377–389. MR3516007

BABANEZHAD, M., VANSTEELANDT, S. and GOETGHEBEUR, E. (2010). Comparison of causal effect estimators under exposure misclassification. *J. Statist. Plann. Inference* **140** 1306–1319. MR2581132

BACCINI, M., MATTEI, A., MEALLI, F., BERTAZZI, P. A. and CARUGNO, M. (2017). Assessing the short term impact of air pollution on mortality: A matching approach. *Environ. Health* **16** 7.

BEELEN, R., RAASCHOU-NIELSEN, O., STAFOGGIA, M., ANDERSEN, Z. J., WEINMAYR, G., HOFFMANN, B., WOLF, K., SAMOLI, E., FISCHER, P. and NIEUWENHUIJSEN, M. (2014). Effects of long-term exposure to air pollution on natural-cause mortality: An analysis of 22 European cohorts within the multicentre ESCAPE project. *Lancet* **383** 785–795.

BICKEL, P. J., GÖTZE, F. and VAN ZWET, W. R. (2012). Resampling fewer than n observations: Gains, losses, and remedies for losses. In *Selected Works of Willem van Zwet* 267–297. Springer, Berlin.

BRAUN, D., GORFINE, M., PARMIGIANI, G., ARVOLD, N. D., DOMINICI, F. and ZIGLER, C. (2017). Propensity scores with misclassified treatment assignment: A likelihood-based adjustment. *Biostatistics* **18** 695–710. MR3799597

BURTON, R. M., SUH, H. H. and KOUTRAKIS, P. (1996). Spatial variation in particulate concentrations within metropolitan Philadelphia. *Environ. Sci. Technol.* **30** 400–407.

CARROLL, R. J. and STEFANSKI, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Amer. Statist. Assoc.* **85** 652–663. MR1138349

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York; CRC Press, London. MR0095561

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144

DEHEJIA, R. H. and WAHBA, S. (1998). Propensity score matching methods for non-experimental casual studies.

DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* **94** 1053–1062.

DI, Q., KOUTRAKIS, P. and SCHWARTZ, J. (2016). A hybrid prediction model for $PM_{2.5}$ mass and components using a chemical transport model and land use regression. *Atmos. Environ.* **131** 390–399.

DI, Q., KLOOG, I., KOUTRAKIS, P., LYAPUSTIN, A., WANG, Y. and SCHWARTZ, J. (2016). Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environ. Sci. Technol.* **50** 4712–4721.

DOCKERY, D. W., POPE, C. A., XU, X., SPENGLER, J. D., WARE, J. H., FAY, M. E., FERRIS JR., B. G. and SPEIZER, F. E. (1993). An association between air pollution and mortality in six US cities. *N. Engl. J. Med.* **329** 1753–1759.

DOMINICI, F., ZEGER, S. L. and SAMET, J. M. (2000). A measurement error model for time-series studies of air pollution and mortality. *Biostatistics* **1** 157–175.

EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903

FULLER, W. A. (2009). *Measurement Error Models*. Wiley Interscience, Hoboken, NJ. Reprint of the 1987 original, *Wiley-Interscience Paperback Series*. MR2301581

GAIL, M. H., WU, J., WANG, M. et al. (2016). Calibration and seasonal adjustment for matched case-control studies of vitamin D and cancer. *Stat. Med.* **35** 2133–2148. MR3513504

GRYPARIS, A., PACIOREK, C. J., ZEKA, A., SCHWARTZ, J. and COULL, B. A. (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10** 258–274.

HARDER, V. S., STUART, E. A. and ANTHONY, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol. Methods* **15** 234–249.

HART, J. E., LIAO, X., HONG, B., PUETT, R. C., YANOSKY, J. D., SUH, H., KIOUMOURT-ZOGLOU, M.-A., SPIEGELMAN, D. and LADEN, F. (2015). The association of long-term exposure to $PM_{2.5}$ on all-cause mortality in the Nurses' Health Study and the impact of measurement-error correction. *Environ. Health* **14** 38.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826

IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866. MR2090918

IMBENS, G. W. (2000). The role of the propensity score in estimating dose–response functions. *Biometrika* **87** 706–710. MR1789821

KIOUMOURTZOGLOU, M.-A., SPIEGELMAN, D., SZPIRO, A. A., SHEPPARD, L., KAUF-MAN, J. D., YANOSKY, J. D., WILLIAMS, R., LADEN, F., HONG, B. and SUH, H. (2014). Exposure measurement error in $PM_{2.5}$ health effects studies: A pooled analysis of eight personal exposure validation studies. *Environ. Health* **13** 2.

KIOUMOURTZOGLOU, M.-A., SCHWARTZ, J., JAMES, P., DOMINICI, F. and ZANOBETTI, A. (2016). PM2.5 and mortality in 207 US cities: Modification by temperature and city characteristics. *Epidemiology* **27** 221–227.

LECHNER, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labour Market Policies* 43–58. Springer, Berlin.

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960.

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425.

PIERCE, D. A. and KELLERER, A. M. (2004). Adjusting for covariate errors with nonparametric assessment of the true covariate distribution. *Biometrika* **91** 863–876. MR2126038

RASSEN, J. A., SHELAT, A. A., FRANKLIN, J. M., GLYNN, R. J., SOLOMON, D. H. and SCHNEEWEISS, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* **24** 401–409.

ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using sub-classification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

ROTHMAN, K. J., GREENLAND, S. and LASH, T. L. (2008). *Modern Epidemiology*. Lippincott Williams & Wilkins, Philadelphia, PA.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (1990). Formal mode of statistical inference for causal effects. *J. Statist. Plann. Inference* **25** 279–292.

SARNAT, S. E., KLEIN, M., SARNAT, J. A., FLANDERS, W. D., WALLER, L. A., MULHOLLAND, J. A., RUSSELL, A. G. and TOLBERT, P. E. (2010). An examination of exposure measurement error from air pollutant spatial variability in time-series studies. *J. Expo. Sci. Environ. Epidemiol.* **20** 135–146.

SHI, L., ZANOBETTI, A., KLOOG, I., COULL, B. A., KOUTRAKIS, P., MELLY, S. J. and SCHWARTZ, J. D. (2016). Low-concentration PM2.5 and mortality: Estimating acute and chronic effects in a population-based study. *Environ. Health Perspect.* **124** 46.

SZPIRO, A. A., SHEPPARD, L. and LUMLEY, T. (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics* **12** 610–623.

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

VAN ROOSBROECK, S., LI, R., HOEK, G., LEBRET, E., BRUNEKREEF, B. and SPIEGELMAN, D. (2008). Traffic-related outdoor air pollution and respiratory symptoms in children: The impact of adjustment for exposure measurement error. *Epidemiology* **19** 409–416.

VAUGHN, B. K. (2008). Data analysis using regression and multilevel/hierarchical models, by Gelman, A., & Hill, J. *J. Educ. Meas.* **45** 94–97.

VILLENEUVE, P. J., WEICHENTHAL, S. A., CROUSE, D., MILLER, A. B., TO, T., MARTIN, R. V., VAN DONKELAAR, A., WALL, C. and BURNETT, R. T. (2015). Long-term exposure to fine particulate matter air pollution and mortality among Canadian women. *Epidemiology* **26** 536–545.

WHO (2018). World Health Organization. Air quality guidelines. Available at http://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

WILSON, W. E. and SUH, H. H. (1997). Fine particles and coarse particles: Concentration relationships relevant to epidemiologic studies. *J. Air Waste Manage. Assoc.* **47** 1238–1249.

WU, X., BRAUN, D., KIOUMOURTZOGLOU, M.-A., CHOIRAT, C., DI, Q. and DOMINICI, F. (2019). Supplement to "Causal inference in the context of an error prone exposure: Air pollution and mortality." DOI:10.1214/18-AOAS1206SUPP.

YANG, S., IMBENS, G. W., CUI, Z., FARIES, D. E. and KADZIOLA, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* **72** 1055–1065. MR3591590

X. WU
D. BRAUN
C. CHOIRAT
Q. DI
F. DOMINICI
DEPARTMENT OF BIOSTATISTICS
HARVARD T.H. CHAN SCHOOL OF
   PUBLIC HEALTH
677 HUNTINGTON AVENUE
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: wuxiao@g.harvard.cedu
        dbraun@mail.harvard.edu

M.-A. KIOUMOURTZOGLOU
DEPARTMENT OF ENVIRONMENTAL
   HEALTH SCIENCES
COLUMBIA UNIVERSITY
   MAILMAN SCHOOL OF PUBLIC
   HEALTH
722 W 168TH STREET
NEW YORK, NEW YORK 10032
USA
E-MAIL: mk3961@cumc.columbia.edu