*To be read before The Royal Statistical Society at a meeting organized by the Environmental Statistics Section and Discussion Meetings Committee on Thursday, 12 December 2024, Dr Adam Sykulski in the Chair.*

# Methods for Estimating the Exposure-Response Curve to Inform the New Safety Standards for Fine Particulate Matter

Michael Cork, Daniel Mork, Francesca Dominici

**Abstract**

Exposure to fine particulate matter ($PM_{2.5}$) poses significant health risks and accurately determining the shape of the relationship between $PM_{2.5}$ and health outcomes has crucial policy implications. Although various statistical methods exist to estimate this exposure-response curve (ERC), few studies have compared their performance under plausible data-generating scenarios. This study compares seven commonly used ERC estimators across 72 exposure-response and confounding scenarios via simulation. Additionally, we apply these methods to estimate the ERC between long-term $PM_{2.5}$ exposure and all-cause mortality using data from over 68 million Medicare beneficiaries in the United States. Our simulation indicates that regression methods not placed within a causal inference framework are unsuitable when anticipating heterogeneous exposure effects. Under the setting of a large sample size and unknown ERC functional form, we recommend utilizing causal inference methods that allow for nonlinear ERCs. In our data application, we observe a nonlinear relationship between annual average $PM_{2.5}$ and all-cause mortality in the Medicare population, with a sharp increase in relative mortality at low $PM_{2.5}$ concentrations. Our findings suggest that stricter limits on $PM_{2.5}$ could avert numerous premature deaths. To facilitate the utilization of our results, we provide publicly available, reproducible code on Github for every step of the analysis.

Keywords: air pollution, all-cause mortality, causal inference, exposure-response curve, fine particulate matter, simulation study

# 1 Introduction

In 2019, air pollution contributed to an estimated 6 million deaths worldwide, accounting for nearly 12% of total global mortality (Health Effects Institute, 2020; Murray et al., 2020). As the leading environmental risk factor for premature mortality, air pollution contributes to a higher number of estimated deaths each year than traffic collisions (Health Effects Institute, 2020; Murray et al., 2020; Manisalidis et al., 2020). Long-term exposure to ambient fine particulate matter ($PM_{2.5}$) is the largest driver of the burden of disease from air pollution worldwide (Health Effects Institute, 2020). In the United States, more than 40% of the population resides in counties with unhealthy levels (above 35 $\mu g/m^3$) of particle pollution (Ameircan Lung Association, 2022). Although the detrimental health effects of $PM_{2.5}$ are widely acknowledged, the shape of the relationship between particle pollution and adverse health outcomes, known as the exposure-response curve (ERC), remains uncertain. Accurate characterization of the ERC, particularly at lower exposure levels, has significant policy implications.

In 2021, guided by epidemiological findings on the relationship between fine particulate matter and mortality risk, the World Health Organization (WHO) released new Global Air Quality Guidelines. These guidelines recommend reducing the annual average limit of $PM_{2.5}$ from 10 $\mu g/m^3$ to 5 $\mu g/m^3$ (World Health Organization, 2021). These guidelines are widely used by decision makers and this change will influence the air quality standards implemented by various international governing bodies. In the United States (USA), the Environmental Protection Agency (EPA) relies on the estimated shape of the ERC between air pollution and health outcomes to determine whether to lower the National Ambient Air Quality Standards (NAAQS). The establishment of the NAAQS was predicated on ensuring public health safety within a sufficient "margin of safety," a guideline that is significantly influenced by the form of the ERC (US Environmental Protection Agency, 2015). For example, an ERC illustrating a reduced slope at lower concentrations of $PM_{2.5}$ relative to the average slope observed throughout the entire spectrum of recorded $PM_{2.5}$ levels could steer the determination of the NAAQS threshold towards the point where a notable change in slope is observed. Similarly, identifying a threshold relationship in the ERC can play a crucial role in determining a suitable limit. On the contrary, if the ERC exhibits a sublinear trend, it might require stricter measures to mitigate potential mortality risks.

Epidemiological evidence in the USA has found that health risks persist at $PM_{2.5}$ levels below the current NAAQS set at 12 $\mu g/m^3$ for annual exposure, suggesting the need for a lower standard (Di, Dai, et al., 2017; Wei et al., 2020; Shi et al., 2021; Ward-Caviness et al., 2021; EPA, 2019). In January 2023, the EPA requested public comment on a proposal to lower the NAAQS for annual $PM_{2.5}$ concentrations. The EPA estimates that adopting stricter standards could prevent up to 4,200 premature deaths annually (Davenport, 2023). Recent evidence suggests that the effects of $PM_{2.5}$ can vary among marginalized subpopulations delineated by racial identity (Black vs. White) and income level (Medicaid eligible vs. ineligible), underlining the potential benefits of reduced levels of $PM_{2.5}$ for these communities (Josey et al., 2023). However, a significant challenge in implementing more stringent ambient air quality standards, through setting a lower emission threshold, lies in establishing conclusive evidence of the causal link between particulate matter and health outcomes based on epidemiological data. Estimating the shape of the ERC and quantifying its uncertainty is complicated by the misaligned nature of the data, potential confounding factors, and effect heterogeneity.

To overcome these challenges, various methods have been proposed and implemented to characterize ERCs. In the field of air pollution epidemiology, ERCs are traditionally fit using a multivariate regression model with the health outcome as the dependent variable, air pollution exposure as an independent variable, and many potential confounders as additional independent variables (Di, Wang, et al., 2017; Liu et al., 2019). Past research has often imposed the assumption of linearity between the natural log of the hazard ratio with respect to $PM_{2.5}$, restricting the ERC to represent an exponential increase in mortality per unit increase in pollutant concentration. Nonparametric or nonlinear functional regression methods provide greater flexibility in describing the shape of the ERC but may not adequately capture threshold effects (Nasari et al., 2016). Other developed methods like the Extended Shape Constrained Health Impact Function (SCHIF) allow nonlinear association and can capture threshold behavior but impose more restrictive model specifications (Burnett et al., 2018). Furthermore, some scientists argue against relying solely on regression-based ERC estimation in epidemiological studies, citing a lack of causal evidence, and advocate for the use of causal inference methods to inform air pollution policies (Chartered Clean Air Scientific Advisory Committee, 2018; Goldman et al., 2019). Indeed, the EPA prioritizes studies employing causal inference methods to determine the limits of the NAAQS (Owens et al., 2017).

Causal inference methods are often placed in the potential outcomes framework (Rubin, 1974), which distinguishes between the design and analysis stages (Imbens and Rubin, 2015). In the design stage, researchers define the causal estimands and target population, and employ design-based methods like matching or weighting to construct a data set that emulates an experimental setup, where units with similar characteristics are compared across various exposure scenarios. After evaluating design quality using metrics such as covariate balance, researchers proceed to the analysis stage to estimate causal effects. The field of environmental health has advocated for and incorporated causal inference frameworks in analyzing the impact of environmental exposures on health outcomes (M.-A. Bind, 2019; M.-A. C. Bind et al., 2019; Carone et al., 2020; National Academies of Sciences, Engineering, and Medicine, 2022). Numerous recent studies have employed these frameworks to delineate the effects of air pollution on health, demonstrating the growing importance and application of causal inference in this field (Zigler et al., 2016; Francesca Dominici et al., 2017; National Academies of Sciences, Engineering, and Medicine, 2022; Brewer et al., 2023)

Causal inference methods, under certain assumptions, exhibit greater robustness to model misspecification compared to traditional regression methods (Imbens and Rubin, 2015). However, many causal inference approaches make the simplified assumption of a binary exposure (Robins, 2000; Hernán et al., 2000; van der Laan et al., 2011; Rosembaum et al., 1983; Rubin and Thomas, 1996). Many methods to estimate causal ERCs focused on approaches that use the generalized propensity score (GPS) (Robins, 2000; Hirano et al., 2005; Robins et al., 1994; Imai et al., 2004). These approaches include outcome modeling (Callaway et al., 2020; Imbens, 2004; Zhao et al., 2020), the use of weighting techniques (Ai et al., 2022; C. Fong et al., 2018; Huber et al., 2020; Yiu et al., 2018), and the implementation of machine learning strategies (Kreif et al., 2015; Zhu et al., 2015). However, these methods are sensitive to misspecification of the GPS model and extreme weights. Doubly robust approaches mitigate this issue and provide asymptotically unbiased estimates of the ERC when either the outcome model or GPS model are misspecified (Kennedy et al., 2017; Colangelo et al., 2022; Schulz et al., 2021). Recently, weighting methods that directly op-

timize covariate balance have been extended to the continuous exposure setting (Vegetabile et al., 2021; Tübbicke, 2022; Huling et al., 2023). Another contemporary development focuses on the extension of the matching framework to the context of continuous exposures through a GPS caliper matching framework (Wu, Mealli, et al., 2022). Each proposed method has been tailored to address specific requirements in causal inference. However, limited research has compared the behavior of these methods under different assumed exposure-response relationships and confounding mechanisms.

In this paper, we aim to fill this research gap by conducting a comprehensive simulation study to compare various estimators of the exposure-response curve. We evaluate the performance of seven estimators, including both regression and causal inference methods, in a range of plausible ERC scenarios and confounding relationships. Furthermore, we apply these methods to estimate the ERC between long-term $PM_{2.5}$ exposure levels and all-cause mortality in a large observational administrative cohort comprising more than 68 million Medicare beneficiaries in the continental United States from 2000 to 2016. In particular, this study represents the first application of multiple statistical approaches to a data set that encompasses more than 500 million person-years of Medicare data. To ensure reproducibility and facilitate public access, we utilize methods available on CRAN and provide code for each step of the analysis on Github. The analysis code can be accessed at https://github.com/macork/ERC_simulations, while the code detailing the data processing for our data application can be found at https://github.com/NSAPH/National-Causal-Analysis.

# 2 Methods

## 2.1 Estimand and estimators

We begin by providing a formal definition of our target estimand. Let $E_i \in \mathbb{R}^+$ represent a continuous nonnegative treatment or exposure for the $i^{th}$ unit in our target population. In alignment with the potential outcomes framework (Rubin, 1974) and adopting the notation established by Imbens and Hirano (Imbens, 2000; Imbens, 2004), the random variable $Y_i(e)$ denotes the potential outcome for subject $i$ when exposed to a level $e \in \mathcal{E}$ of exposure, where $\mathcal{E}$ denotes all levels of the exposure. We are interested in the random variable $Y_i(e)$, which represents the potential outcome at different levels of exposure. However, note that $Y_i(e)$ cannot be directly estimated between various exposure values since $Y_i(e)$ is observed only for one exposure value for each unit, known as the fundamental problem of causal inference. Therefore, we shift our target estimand to the population ERC $R(e)$, defined as the expected potential outcome at each exposure level $R(e) = E[Y_i(e)]$, where the expectation is taken over the distribution of the counterfactual result in the population of interest.

In most scenarios, treatments or exposures are not randomly allocated across the population of interest, even within experimental settings. Consequently, to achieve consistent estimates of the ERC in observational studies, certain identifiability assumptions are required. First, given the observed pretreatment covariates $X$, the potential outcomes must be independent of the received treatment or exposure.

$$Y_i(e) \perp\!\!\!\perp E_i \mid X_i \quad \forall e \in \mathcal{E} \tag{1}$$

This principle, often referred to as unconfoundedness, selection on observables (Heckman et al., 1985), or the conditional independence assumption (Lechner, 2001), presupposes that the researcher has identified and measured all covariates that influence both the assignment to exposure and the potential outcomes. The next assumption establishes the need for a common support. Specifically, the conditional density of the treatment or exposure should maintain a positive value across $\mathcal{E}$.

$$f_{E|X}\left(E = e \mid X_i\right) > 0 \quad \forall e \in \mathcal{E} \tag{2}$$

Compared to a binary exposure scenario, this condition can be considerably more stringent, especially in areas with limited exposure density. Under such conditions, trimming the sample and estimating the ERC from a subset of observations may be necessary to avoid extrapolation (Crump et al., 2009; Lechner and Strittmatter, 2019). Finally, we adopt the Stable Unit Treatment Value Assumption (SUTVA), stating that an individual's outcome depends exclusively on their specific exposure (Rubin, 1980), thus ruling out spillover effects. Although these assumptions are frequently invoked to identify the ERC in observational studies, they might not be sufficient to ensure a particular estimator's accurate discernment of the ERC.

In this analysis, we compared the performance of seven different estimators in approximating $R(e)$ across various assumed exposure-response relationships and confounding mechanisms (Table 1). The first three estimators used to approximate the ERC are regression methods often used in practice. These regression methods included a linear model that adjusts for all variables in a linear manner (linear), a generalized additive model that accounts for confounders linearly while allowing for a flexible nonlinear relationship between exposure and outcome through a spline basis representation (GAM), and a change point model that fits a segmented linear model with a single break point (change point). These traditional regression methods do not separate between a design and analysis stages. In contrast, the remaining four estimators are design-based methods commonly used in the field of causal inference. Specifically, we employed the same linear, GAM, and change point outcome models. However, instead of adjusting for confounding using linear regression terms, we utilized entropy balancing to generate continuous treatment weights. Entropy balancing is a method to create optimal weights to balance covariate distributions across exposure levels, thus flexibly controlling for observed confounders in causal estimation (Hainmueller, 2012). This method works by solving a convex optimization problem, with the aim of minimizing deviations from uniform base weights while ensuring zero correlation and adhering to normalization constraints (Tübbicke, 2022; Vegetabile et al., 2021). In its simplest form for a continuous exposure, entropy balancing generates weights that eliminate the correlation between the exposure variable and all covariates. However, removing the correlation alone may not be enough to ensure independence. Therefore, entropy balancing can be expanded to include higher moments of the treatment variable (Yiu et al., 2018; Tübbicke, 2022). To mitigate the presence of extreme weights, we generated weights and then truncated the upper 0.5% of weights to the 99.5% percentile. Once we obtained the weights, we estimated the ERC using the same regression techniques based on the weighted sample (linear entropy, GAM entropy, and change point entropy). The seventh and final estimator, CausalGPS, extends the matching to the context of continuous exposure through a GPS

caliper matching framework (Wu, Mealli, et al., 2022). Further details on this estimator can be found in the CausalGPS package and the supplementary materials.

| Estimator name | Description | Algorithm | Software package |
|---|---|---|---|
| **Linear** | Gaussian linear model that linearly adjusts for expsoure ($E$) and covariates ($\mathbf{C}$) | Chapter 4 (Chambers et al., 1992) | stats (R Core Team, 2021) |
| **GAM** | Generalized linear model employing spline with four degrees of freedom to adjust for $E$ and linearly adjusts for $\mathbf{C}$ | Chapter 3, 4 (Wood, 2017) | mgcv (Wood, 2017) |
| **Change point** | Segmented threshold regression model with threshold value based on $E$, model linearly adjusts for $\mathbf{C}$ | Implementation section (Y. Fong et al., 2017) | chngpt (Y. Fong et al., 2017) |
| **Linear entropy** | Same outcome model as **linear** with entropy-based weight adjustment | Section 3 (Tübbicke, 2022) | WeightIt (Greifer, 2022) |
| **GAM entropy** | Same outcome model as **GAM** with entropy-based weight adjustment | Section 3 (Tübbicke, 2022) | WeightIt (Greifer, 2022) |
| **Change point entropy** | Same outcome model as **change point** with entropy-based weight adjustment | Section 3 (Tübbicke, 2022) | WeightIt (Greifer, 2022) |
| **GPS matching (CausalGPS)** | Causal inference method that uses GPS matching in a continuous setting | Section 3.1 (Wu, Mealli, et al., 2022) | CausalGPS (Wu, Mealli, et al., 2022) |

Table 1: **Overview of Estimators for Analyzing the Exposure-Response Curve.** The Algorithm column cites the precise sections within a resource where the methodologies for each estimator are elaborated. The Software package column identifies the R packages employed for implementing each estimator in this study.

## 2.2 Simulation setup

We conducted a comprehensive set of simulations to evaluate the performance of our methods in estimating the effect of a continuous exposure on a continuous outcome under various data-generating mechanisms. In particular, we examine how each estimator performed in different scenarios: (1) when the marginal relationship between exposure and outcome changed, (2) when the relationship between the confounders and exposure shifted, (3) when there was an interaction between the exposure and confounders in the outcome model, and (4) when the sample size varied.

### 2.2.1 Simulation settings

For each observation $i = 1, \ldots, n$ we generated a vector of six covariates $\mathbf{C}_i = (C_{1i}, C_{2i}, ..., C_{6i})$ with five continuous components and one categorical component:

$$(C_{1i}, ..., C_{4i})' \sim \mathcal{MVN}(\mathbf{0}, I_4), C_{5i} \sim V\{-2, 2\}, C_{6i} \sim U(-3, 3)$$

where $\mathcal{MVN}(0, I_4)$ denotes a multivariate normal distribution, $I_4$ is the identity matrix, $V\{-2, 2\}$ denotes a discrete uniform distribution, and $U(-3, 3)$ denotes a continuous uniform distribution. To generate exposure $E_i$, we outlined four specifications of the relationship between confounders and exposure, which based on the cardinal function $\gamma(\mathbf{C}) = -0.8 + (0.1, 0.1, -0.1, 0.2, 0.1, 0.1)\mathbf{C}$. The coefficients of the cardinal function $\gamma(\mathbf{C})$ were similar to those used in previous research (Kennedy et al., 2017; Wu, Mealli, et al., 2022). The relationship between the covariates and the exposure (exposure model) varied in terms of error distribution and complexity, as described by the following equations.

$$\begin{aligned}
E_{\text{linear}} &= 9 * \gamma(\mathbf{C}) + 18 + N(0, 10) \\
E_{\text{heavy-tail}} &= 9 * \gamma(\mathbf{C}) + 18 + \sqrt{5} \times T(3) \\
E_{\text{nonlinear}} &= 9 * \gamma(\mathbf{C}) + 2 * C_3^2 + 15 + N(0, 10) \\
E_{\text{interaction}} &= 9 * \gamma(\mathbf{C}) + 2 * C_3^2 + 2 * C_1 C_4 + 15 + N(0, 10)
\end{aligned}$$

In the first scenario, denoted as $E_{linear}$, a linear relationship between $E$ and $\mathbf{C}$ is assumed, with the addition of normal noise characterized by $N(0, 10)$, representing a normal distribution with a variance of 10. In the $E_{heavy-tail}$ scenario, the relationship between $E$ and $\mathbf{C}$ remains linear, but the exposure values were heavy-tailed and include extreme values. Specifically, we use $T(3)$ to represent a Student's t-distribution with 3 degrees of freedom. The $E_{nonlinear}$ scenario introduced a nonlinear relationship between $E$ and $\mathbf{C}$ while retaining normal noise. In the fourth scenario, $E_{interaction}$, an interaction term was added to the relationship between exposure and confounders. The scale (9) and location (18, 18, 15, 15) parameters were specified to ensure that more than 95% of the exposure values exist between the range of [0, 20]. These simulation scenarios were designed to be comparable to the exposure range observed in our data application. As a final step, we enforced that all exposure values be nonnegative to reflect plausible air pollution exposure values and ensure the validity of all the data-generating mechanisms. This was achieved by oversampling the number of exposure values, allowing us to remove any negative values and then randomly sample the remaining exposure values to match the desired sample size.

In all scenarios, we generated the outcome variable $Y_i$ from a normal distribution $Y|E, C \sim N(\mu(E, C), 10^2)$, where the mean function $\mu(E, C)$ is determined by the exposure ($E$) and the confounders ($C$), and the standard deviation $\sigma$ is set to 10. We considered three different specifications of the exposure-response curves: linear, sublinear, and threshold, denoted $\mu_{\text{linear}}$, $\mu_{\text{sublinear}}$, and $\mu_{\text{threshold}}$, respectively. These three ERCs represent plausible relationships between air pollution and adverse health outcomes. For each specification of the ERC, we examined two outcome

models. The first model, denoted $\mu_{\text{ERC}}$, assumed a linear relationship between confounders and the result, independent of exposure. The second model, denoted $\mu_{\text{ERC,int}}$, incorporated an interaction term between the confounders and the exposure. This interaction term accounts for the presence of heterogeneous treatment effects, where the impact of the exposure on the outcome is influenced by an individual's covariate values. The formulations of the mean functions are provided as follows:

$$
\begin{aligned}
\mu_{\text{linear}}(E, C) &= 20 - (2, 2, 3, -1, 2, 2) * \mathbf{C} + E \\
\mu_{\text{sublinear}}(E, C) &= 20 - (2, 2, 3, -1, 2, 2) * \mathbf{C} + 5\log(E + 1) \\
\mu_{\text{threshold}}(E, C) &= 20 - (2, 2, 3, -1, 2, 2) * \mathbf{C} + 1.5E[E > 5] \\
\mu_{\text{linear, int}}(E, C) &= \mu_{\text{linear}} + E(-0.1C_1 + 0.1C_3^2 + 0.1C_4 + 0.1C_5) \\
\mu_{\text{sublinear, int}}(E, C) &= \mu_{\text{sublinear}} + 3\log(E + 1)(-0.1C_1 + 0.1C_3^2 + 0.1C_4 + 0.1C_5) \\
\mu_{\text{threshold, int}}(E, C) &= \mu_{\text{threshold}} + 1.5E[E > 5](-0.1C_1 + 0.1C_3^2 + 0.1C_4 + 0.1C_5)
\end{aligned}
$$

In total, we generated data for 72 different scenarios combining four exposure specifications, six mean function specifications for our outcome model, and three sample sizes (N = 200, 1000, 10,000). For each scenario, we conducted 1000 simulations, where we simulated the exposure and outcome models in each iteration. We also conducted an additional data-informed simulation where we sampled from the data set used in our application to provide further evidence that our findings are applicable in real-world settings. Refer to Section 2.3.3 for more details.

### 2.2.2 Evaluation of estimators

Under every data-generating scenario we estimated the ERC using all seven estimators. We evaluated the goodness of fit of the estimators by comparing the absolute bias and the root mean squared error (RMSE) of the estimated ERCs. We estimated these two quantities empirically by first evaluating the metric at equally spaced values in the range $\hat{\mathcal{E}}^*$ for all simulation replicates. We then averaged the metrics for each point across the exposure range $\hat{\mathcal{E}}^*$. $\hat{\mathcal{E}}^*$ is a restricted support of $\hat{\mathcal{E}}$ where we excluded some mass at the boundary to avoid boundary instability and only evaluated the ERC from zero to twenty (Kennedy et al., 2017). The two quantities are formally defined as:

$$
\begin{aligned}
|\,\text{Bias}\,| &= M^{-1} \sum_{m=1}^{M} \left| S^{-1} \sum_{s=1}^{S} \hat{R}_s(e_m) - R(e_m) \right| \\
\text{RMSE} &= M^{-1} \sum_{m=1}^{M} \left( S^{-1} \sum_{s=1}^{S} \left( \hat{R}_s(e_m) - R(e_m) \right)^2 \right)^{1/2}
\end{aligned}
\tag{3}
$$

Here, $M = 100$, and $e_1, ..., e_M$ are equally spaced points in the restricted support $\hat{\mathcal{E}}^*$. $\hat{R}_s$ represents the estimate of the ERC in simulation $s$. We report bias and RMSE for each approach averaged across $S = 1000$ simulations.

For causal estimators, we evaluated the quality of the design setup by assessing the covariate balance. Covariate balance measures the similarity of the distribution of observed preexposure covariates across all exposure levels to avoid confounder bias. Previous literature suggests that achieving an absolute correlation of 0.1 or lower indicates empirical covariate balance (Zhu et al., 2015). In the case of entropy balancing, we calculated the absolute correlation by considering the weighted correlation between each covariate and the exposure, using the weights generated during the design phase. The CausalGPS method follows a similar procedure to compute the measure of absolute correlation (Wu, Mealli, et al., 2022).

## 2.3 Data application

### 2.3.1 Data set

We apply the proposed methods to estimate the effect of long-term $PM_{2.5}$ exposure on all-cause mortality using a previously identified data set (Josey et al., 2023). We used the Medicare enrollee cohort in the contiguous USA from 2000 to 2016, which includes demographic information such as age, sex, race/ethnicity, date of death, and residential ZIP code. The study population consisted of over 68 million individuals residing in 31,414 ZIP codes, for which we compiled the number of deaths among Medicare enrollees for each ZIP code and year. Annual estimates of $PM_{2.5}$ exposure were obtained from a validated ensemble prediction model developed in previous research (Di, Amini, et al., 2019). To obtain the annual average $PM_{2.5}$ in each ZIP code, we aggregated the daily $PM_{2.5}$ exposure estimates at a 1km x 1km grid cell resolution using area-weighted averages (Di, Wang, et al., 2017). We assigned the annual average $PM_{2.5}$ value to individuals residing in each ZIP code for each calendar year. The predicted annual average $PM_{2.5}$ ranged from 0.01 to 30.92 $\mu g/m^3$, with the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentiles being 4.26 and 15.04, respectively. To avoid extrapolation at the limits of the exposure range, we trimmed the highest 5% and lowest 5% of $PM_{2.5}$ exposures in the ERC. This trimming practice aligns with previous literature (Liu et al., 2019; Di, Wang, et al., 2017) and assists in meeting the causal assumptions necessary to identify our estimand of interest (Petersen et al., 2012).

To address potential confounding factors, the data set incorporated 10 variables at the ZIP code and county level. These variables included ZIP code-level socioeconomic status indicators (SES) obtained from the 2000 and 2010 Census and the 2005-2012 American Community Surveys (ACS), as well as county-level information from the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS) (Wu, Braun, et al., 2020; Josey et al., 2023). Specifically, potential confounders consisted of two county-level variables: average body mass index and smoking rate; and eight census variables at the ZIP code level: proportion of Hispanic residents, proportion of Black residents, median household income, median home value, proportion of residents in poverty, proportion of residents with a high school diploma, population density, and proportion of residents who own their home. Furthermore, we include four meteorological variables at the ZIP code level: summer (June to September) and winter (December

to February) averages of maximum daily temperatures and relative humidity. The data set also included two indicator variables indicating (i) the four census geographic regions of the United States (Northeast, South, Midwest, and West) and (ii) calendar years (2000-2016) to adjust, respectively, for any residual or unmeasured spatial and temporal confounding.

#### 2.3.2 Analysis

We used each of the methods described in our simulation to estimate the ERC between $PM_{2.5}$ and all-cause mortality. The outcome variable used was the log-transformed mortality rate, while the covariates at the ZIP code level, along with the year and strata variables, were included in the models. To address zero rates, we replaced them with half the minimum observed mortality rate across all ZIP codes and years. Following the proposed methodology, we fit each model and exponentiated the mean estimates to obtain the estimated all-cause mortality rate as a function of $PM_{2.5}$ concentration. The entire ERC was reconstructed using estimates at 100 equidistant levels of exposure, ranging from the minimum to the maximum observed $PM_{2.5}$ concentration in our sample.

We aggregated the data set in this study at the ZIP code-year level and then stratified based on age, sex, Medicaid eligibility (as a proxy for individual-level socioeconomic status) and follow-up year. Since each ZIP code-year contained varying amounts of person-time contributing to each stratum, we employed a weighted regression with person-time as weights to accurately estimate the population ERC. However, we adapted the general approach for our causal inference estimators due to the misalignment between the exposure and outcome measurements. Exposure and confounder data were measured at the ZIP code-year level, whereas mortality data incorporate demographic and structural strata within ZIP code-years. To address this discrepancy, we modified our approach, drawing on previous work (Josey et al., 2023; Balzer et al., 2019). First, to account for confounding, we generated weights (entropy balancing weights for entropy balancing estimators and "matching weights" for CausalGPS) that balanced the distribution of covariates across different exposure levels at the ZIP code-year level. Note that for entropy balancing weights, we required that the first and second moments of exposure $PM_{2.5}$ be uncorrelated with all covariates. This mitigated confounding effects by breaking the association between exposure and covariates in the weighted data. Second, we specified an outcome model in which we modeled mortality rates specific to ZIP code and year as a function of both individual-level and ZIP code-level covariates. Finally, we fit the outcome model by multiplying the balancing weights from the design stage of our analysis with the observed person-time within each ZIP code-year stratum. The validity of this weighting procedure has been outlined in the literature (Dong et al., 2020; Zanutto, 2006). Further details regarding each step can be found in the supplementary materials.

We implemented the M-out-of-N bootstrap procedure, where M = N/log(N), to construct the point-wise Wald 95% confidence band for the ERC (Politis et al., 1994; Bickel et al., 2012). In this procedure, we utilized a block bootstrap with ZIP codes serving as block units. This approach allowed us to consider the correlation between observations across different years but within the same ZIP code (Wu, Mealli, et al., 2022). For each bootstrap replicate, we recalculated the GPS and entropy weights and refit the outcome model. This ensured that the bootstrap procedure accounted for the variability associated with both the design and analysis stages of the causal inference estimators.

#### 2.3.3 Data-informed simulation

We carried out an additional simulation exercise using Medicare data to emulate key attributes of the data set used in our main application, with the aim of providing additional evidence that the findings of our simulation study are applicable in empirical practice. In this simulation, we selected a random sample from the Medicare data set (N = 10,000), with each observation sampled according to zip code, year, age group, gender, race, and Medicaid eligibility status. We then incorporated the observed annual average $PM_{2.5}$ concentration along with six covariates from the Medicare data: Mean BMI, Median Home Value, Percentage of Individuals Who Have Ever Smoked, Percentage with Below High School Education, Percentage Below the Poverty Level, and Average Winter Temperature. Subsequently, simulations were conducted based on the six mean models described in Section 2.2.1. The covariates were standardized to have a mean of 0 and a standard deviation of 1 to ensure the outcome was within a plausible range.

# 3 Results

## 3.1 Simulation results

### 3.1.1 Covariate balance

Before assessing the results of our simulation, we first evaluated the performance of our causal inference estimators in achieving covariate balance. Covariate balance is determined by the relationship between the covariates and the exposure, and it should be similar across all outcome models in our simulation. Good covariate balance is demonstrated when the average absolute correlation between the covariates and the exposure is below 0.1 in the reweighted data set. In Supplementary Figure S1, we present the number of simulations that achieved covariate balance using either entropy balancing weights or the CausalGPS estimator. Entropy balancing weights consistently achieved covariate balance, regardless of the sample size, as they are designed to limit the absolute correlation. They only failed to achieve covariate balance in the heavy-tailed exposure setting under large sample sizes, which can be attributed to our truncation procedure. The GPS matching framework used in the CausalGPS estimator tended to achieve covariate balance in larger sample sizes (N = 10,000) or in the linear and heavy-tailed exposure models. On the other hand, it often failed to achieve balance in the nonlinear or interaction exposure scenarios at lower sample sizes. For example, with a sample size of 200, only 26% of simulations across all exposure models had a mean absolute correlation below 0.1 for the CausalGPS estimator, compared to 67% and 93% with sample sizes of 1,000 and 10,000, respectively. Additionally, in Supplementary Figure S2, we report the average and upper and lower bounds for the correlation achieved for each individual covariate using both entropy weights and the CausalGPS package, based on a sample size of 1,000. Note that a mean absolute correlation below 0.1 does not guarantee that the absolute correlation falls below 0.1 for each individual covariate.

### 3.1.2 Linear outcome model

We begin by presenting our results for the linear ERC scenario ($\mu_{\text{linear}}, \mu_{\text{linear,int}}$). Figure 1 compares the absolute bias and RMSE of the seven estimators for a sample size of 1,000. Supplementary Figures S3 and S4 provide the absolute bias and RMSE across all sample sizes. In the absence of an interaction term ($\mu_{linear}$), all estimators exhibited relatively low bias. The linear model demonstrated the least bias and RMSE across different confounder settings. On the other hand, the CausalGPS estimator showed the highest bias and RMSE, which decreased with increasing sample size (Supplementary Figure S3).

In contrast, under an interaction outcome setting ($\mu_{linear,int}$), we observed that causal inference methods generally outperformed regression methods in terms of absolute bias. The entropy weights effectively debiased the results in the linear and heavy-tailed exposure settings. However, the use of entropy balancing weights resulted in an increase in absolute bias when exposure was defined by a nonlinear function of the covariates ($E_{\text{nonlinear}}, E_{\text{interaction}}$). In the nonlinear and interaction exposure settings, the CausalGPS estimator exhibited lower absolute bias compared to entropy-weighted or non-causal methods. Regarding RMSE, the causal inference methods generally yielded similar results to regression methods, except for the CausalGPS estimator, which showed higher RMSE. However, the RMSE of the CausalGPS method reached a level comparable to that of other estimators at larger sample sizes (Supplementary Figure S4). Supplementary Figure S5 displays the ERCs in the $\mu_{\text{linear,int}}$ model setting. These plots reveal that while the CausalGPS estimator exhibited minimal bias in its mean ERC curve for the nonlinear and interaction exposure setting, it displayed a higher level of variability. Additionally, the GAM and change point models underestimated the response at higher exposure values.

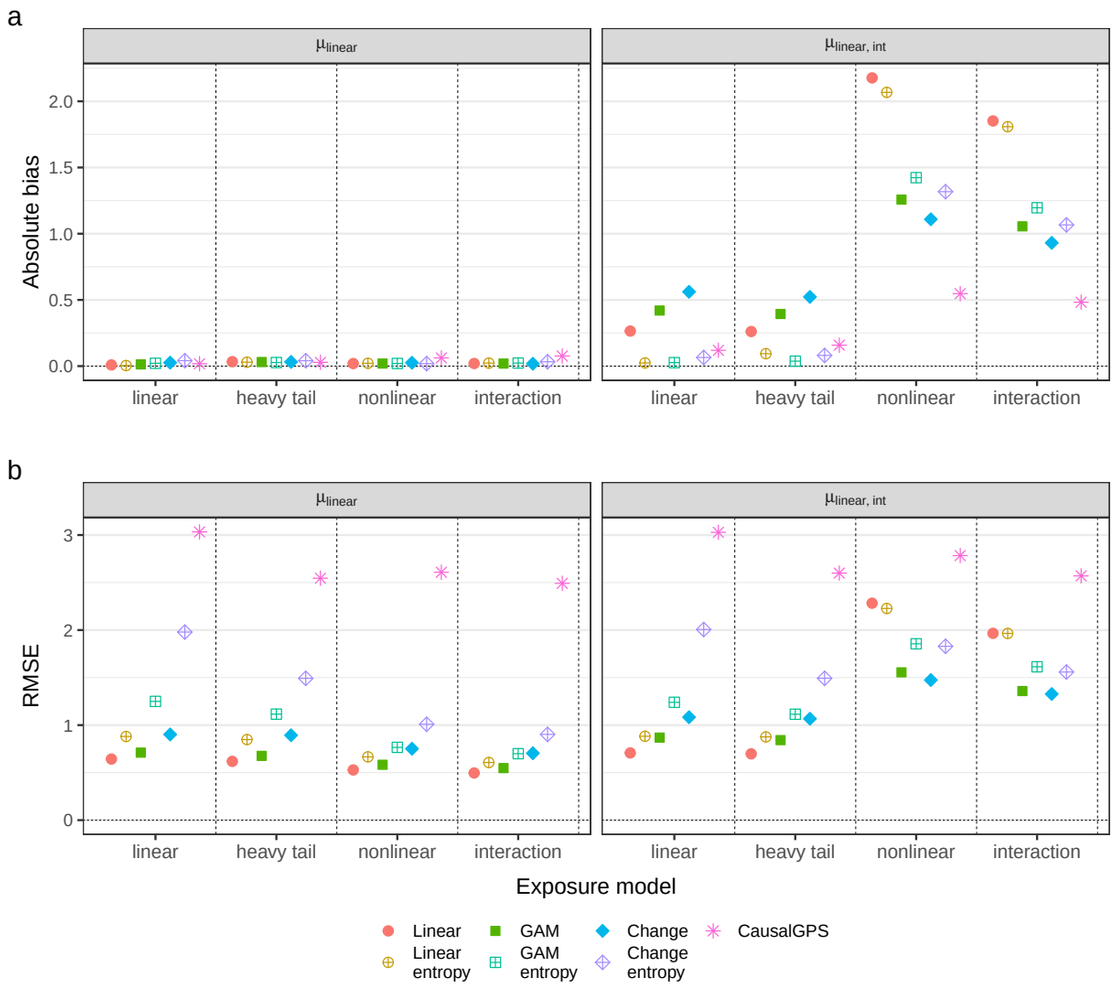


Figure 1: **Absolute bias and RMSE of ERC estimators in the linear outcome model setting**. (a) Absolute bias and (b) RMSE from simulations under different specifications of our estimator with sample size of 1000. We plot the mean absolute bias and RMSE from 1000 simulations. Plots are faceted by the presence or absence of an interaction in the linear outcome model.

### 3.1.3 Sublinear outcome model

Next, we present results under the sublinear outcome model settings ($\mu_{\text{sublinear}}, \mu_{\text{sublinear,int}}$). In Figure 2 we compare the absolute bias and RMSE for the seven estimators with a sample size of 1000. In the sublinear outcome model setting without interaction ($\mu_{sublinear}$), the GAM, change point, and CausalGPS estimators demonstrated the least absolute bias. The GAM models consistently showed the lowest RMSE in this setting, while CausalGPS exhibited a large RMSE.

When considering the presence of heterogeneous treatment effects ($\mu_{sublinear,int}$), the GAM, change point, and CausalGPS estimators again resulted in the least absolute bias in the linear and heavy-tail exposure settings. In the nonlinear and interaction exposure settings, entropy weighting did not debias the ERC, and only the CausalGPS estimator exhibited low bias. The CausalGPS estimator tended to debias more effectively at larger sample sizes (Supple-

mentary Figure S6). Regarding the RMSE, we also observed that the GAM model consistently had the lowest RMSE in the $\mu_{\text{sublinear,int}}$ setting. This result is surprising given the known bias in the model. We observed that the CausalGPS estimator had a larger RMSE, which decreased with the sample size (Supplementary Figure S7). Supplementary Figure S8 demonstrates that only CausalGPS captured the ERC in the nonlinear and interaction exposure setting when assessing the mean ERC at low or high exposure levels.

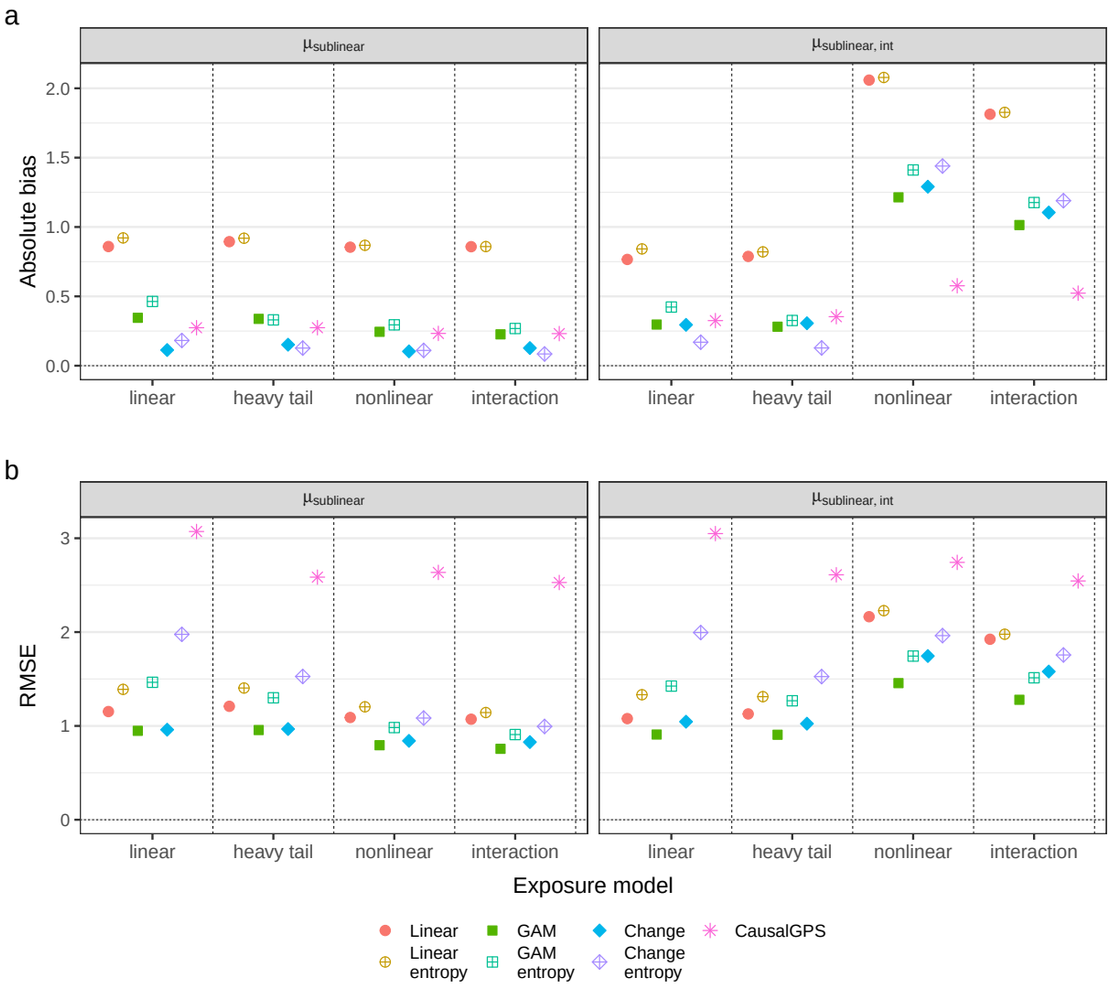


Figure 2: **Absolute bias and RMSE of ERC estimators in the sublinear outcome model setting**. (a) Absolute bias and (b) RMSE from simulations under different specifications of our estimator with sample size of 1000. We plot the mean absolute bias and RMSE from 1000 simulations. Plots are faceted by the presence or absence of an interaction in the sublinear outcome model.

### 3.1.4 Threshold outcome model

We computed the same metrics in the threshold outcome model setting (Figure 3). In the absence of interaction ($\mu_{\text{threshold}}$), we observed that the change point models, with and without entropy weighting, exhibited the least absolute bias. When assessing RMSE, both the change point estimators and the GAM estimators performed well. At larger sample sizes, the change point entropy models showed comparable RMSE (Supplementary Figure S10).

Under an interaction mean outcome model ($\mu_{\text{threshold,int}}$), the change point entropy model demonstrated the lowest absolute bias in the linear or heavy-tailed exposure setting. The CausalGPS estimator exhibited low absolute bias in all exposure settings. When comparing RMSE, the entropy balancing scheme generally led to increased RMSE compared to the unweighted estimators. However, this disparity disappeared at larger sample sizes (Supplementary Figure S10). The GAM and change point models tended to have the lowest RMSE across the four exposure scenarios. When assessing the ERCs in Supplementary Figure S11, few estimators captured the threshold behavior. CausalGPS did not capture an effect below 5 on average in the $E_{nonlinear}$ and $E_{interaction}$ settings and did not attenuate effects at higher exposure levels. The change point entropy estimator captured the threshold only in the $E_{linear}$ and $E_{heavytail}$ settings.

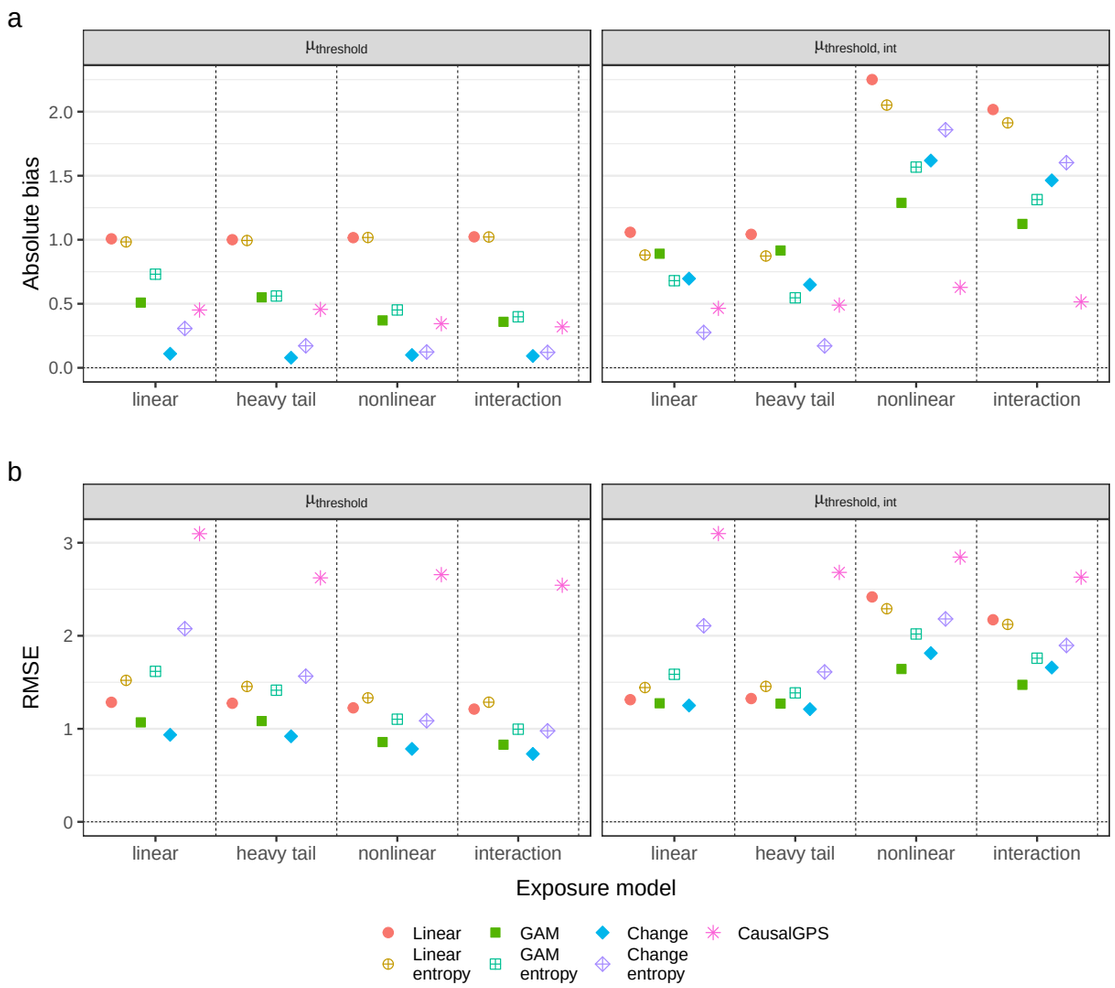


Figure 3: **Absolute bias and RMSE of ERC estimators in the threshold outcome model setting**. (a) Absolute bias and (b) RMSE from simulations under different specifications of our estimator with sample size of 1000. We plot the mean absolute bias and RMSE from 1000 simulations. Plots are faceted by the presence or absence of an interaction in the threshold outcome model.

### 3.1.5 Data-informed simulation

Additional simulation results using Medicare data for exposure and six covariates under the six mean model specifications are presented in Figure 4. For a linear mean outcome model without an interaction term ($\mu_{linear}$), all estimators showed minimal bias. The linear model had the lowest bias and RMSE across various confounder scenarios. However, in an interaction outcome model setting ($\mu_{linear,int}$), causal inference methods outperformed regression methods in absolute bias. The RMSE for causal inference methods was comparable to that of traditional regression methods in this sample size (N = 10,000).

In the non-interaction sublinear outcome model scenario ($\mu_{sublinear}$), the GAM, change point, and CausalGPS estimators had the smallest absolute bias. GAM models resulted in the lowest RMSE, while CausalGPS showed a significant RMSE. In scenarios with heterogeneous treatment effects ($\mu_{sublinear,int}$), change point estimators had the least absolute bias. Notably, causal inference methods could not effectively reduce bias in this context. Regarding RMSE, GAM models consistently presented the lowest RMSE in the $\mu_{\text{sublinear,int}}$ scenario.

In a threshold setting without interaction ($\mu_{\text{threshold}}$), the change point estimators, both with and without entropy weighting, showed the least absolute bias. Under an interaction mean outcome model ($\mu_{\text{threshold,int}}$), the change point entropy estimator showed the lowest absolute bias. The change point estimator had one of the highest RMSE among all evaluated methods across all mean model scenarios, including the ($\mu_{\text{threshold,int}}$) setting. In this particular scenario, causal inference methods outperform traditional regression methods in terms of bias and RMSE.

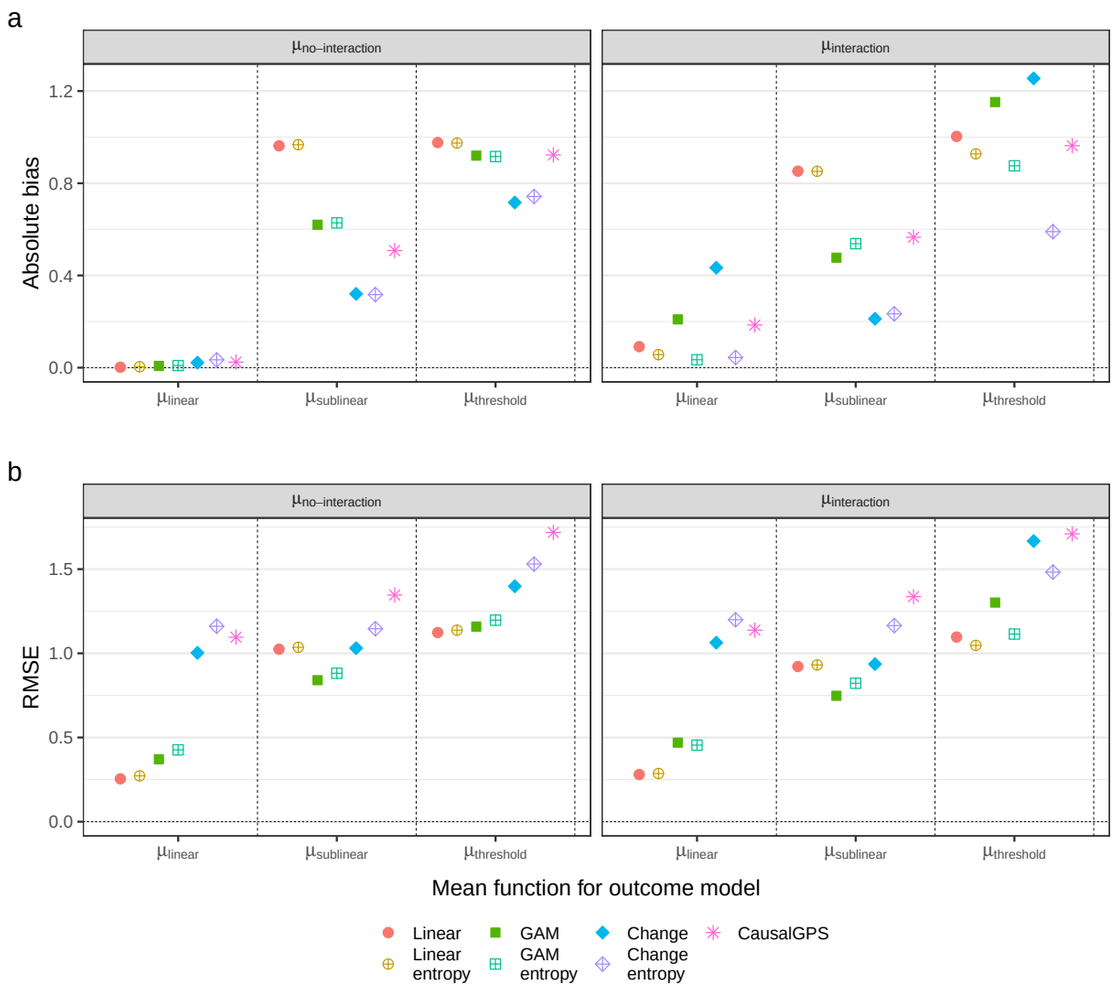


Figure 4: **Absolute bias and RMSE of ERC estimators in all outcome model setting under data-informed simulation setting**. (a) Absolute bias and (b) RMSE of estimators from simulations under different specifications of the mean function of the outcome model with a sample size of 10,000. We plot the mean absolute bias and RMSE from 1000 simulations. The x-axis denotes outcome model type, and the plots are faceted by the presence or absence of an interaction in the mean model.

## 3.2 Data application

We present the covariate balance graphs for the causal inference estimators using the Medicare data in Figure 5. Note that while the absolute correlation between each covariate and exposure was not below 0.1 for the CausalGPS estimator, the mean absolute correlation fell below 0.1.

Figure 6 shows the estimated ERC for mortality rate as a function of the annual average $PM_{2.5}$. The change point models, which failed to converge for this large data set, were omitted from the plot. Figure 6 shows that nonlinear causal inference methods (GAM entropy, CausalGPS) exhibited a more pronounced increase in mortality at lower levels of annual average $PM_{2.5}$ before attenuating at higher concentrations, similar to our sublinear outcome model in our simulations.

Figure 7 presents the estimated relative mortality rates for the Medicare population. We compared the estimated mortality rate for each estimator with the estimated mortality rate at the current EPA limit of 12 $\mu g/m^3$. All estimators projected a decrease in relative mortality for the Medicare population at concentrations lower than 12 $\mu g/m^3$. The CausalGPS estimator demonstrated a more gradual decrease in relative mortality from 9-12 $\mu g/m^3$ compared to other estimators, followed by a sharp decrease at concentrations lower than 9 $\mu g/m^3$. The GAM, GAM entropy, linear, and linear entropy estimators illustrated similar decreases with overlapping confidence intervals.

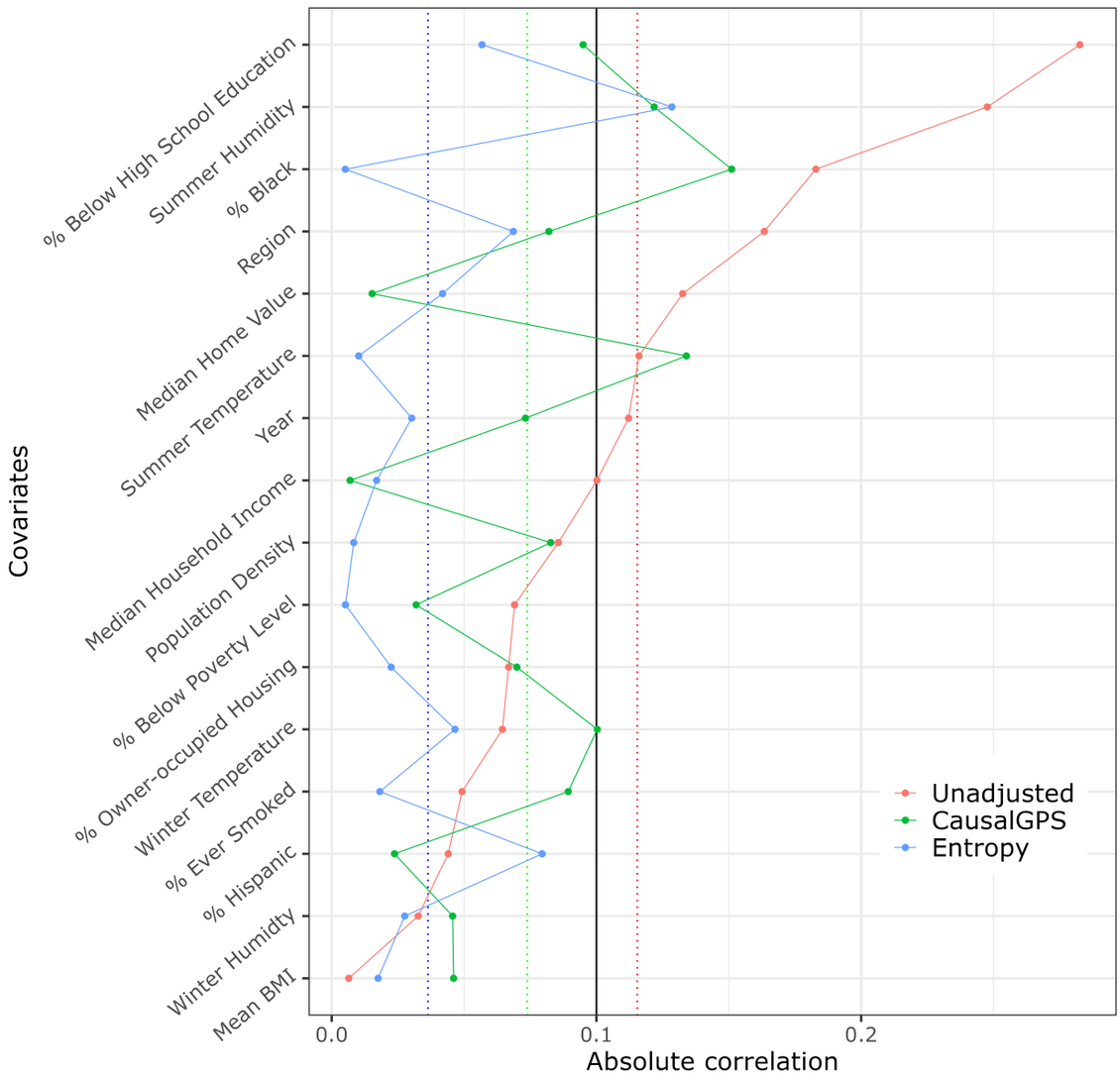


Figure 5: **Covariate balance plots of the absolute Pearson correlation coefficients in the unweighted Medicare data (unadjusted) and after weighing**. Absolute Pearson correlation between each covariates and $PM_{2.5}$ unadjusted (red), after CausalGPS matching adjustment (green), and after entropy weighting adjustment (blue). Absolute correlation of 0.1 (solid black line) or lower is considered empirically achieving covariate balance achieved (Zhu et al., 2015). Mean absolute correlation across all covariates is shown by the dotted line for unadjusted (red) CausalGPS (green) and entropy weighting (blue).

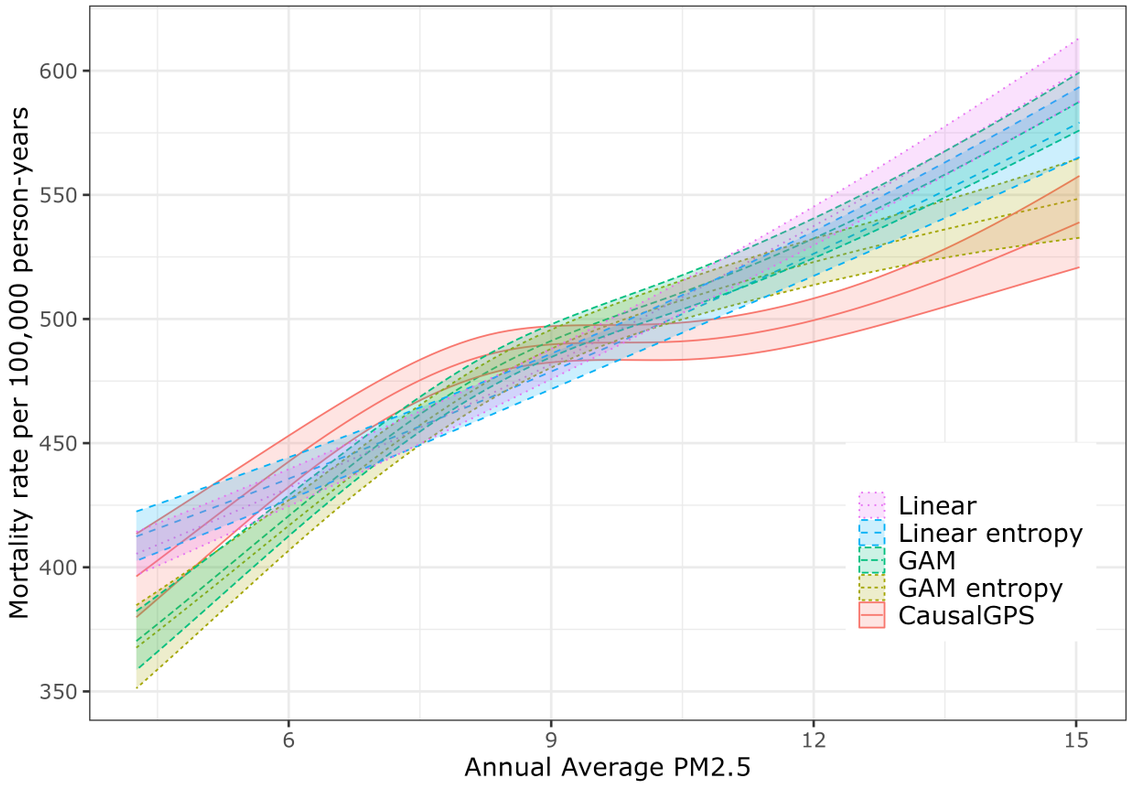


Figure 6: **Estimated ERC relating mortality as a function of annual average $PM_{2.5}$ on Medicare enrollee cohort with 95% confidence intervals**. Estimator are designated by color; models failed to converge for change point and change point entropy models.

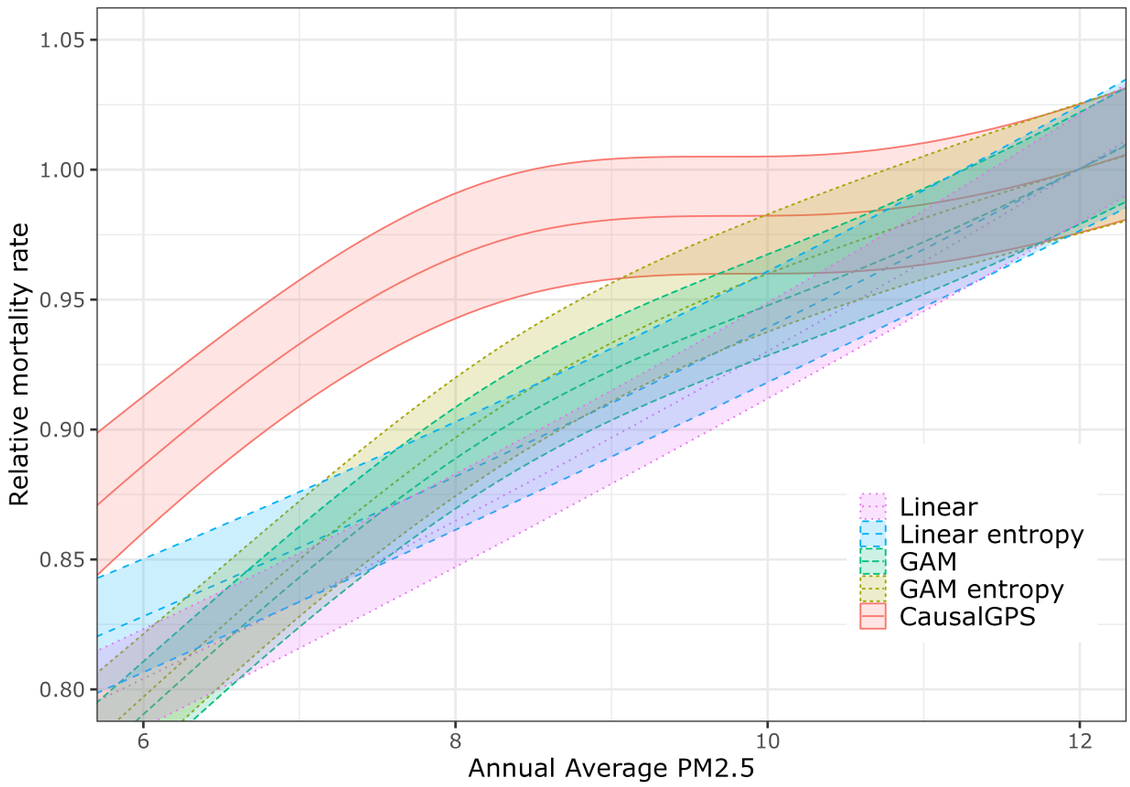


Figure 7: **Point estimates and 95% confidence intervals of the relative mortality rate corresponding to decreases in annual average $PM_{2.5}$ with respect to 12 $\mu g/m^3$ on average for the Medicare population**. Estimator are designated by color; models failed to converge for change point and change point entropy models.

# 4 Discussion

Air pollution remains a significant threat to human health. Determining the shape of the exposure-response curve (ERC) between $PM_{2.5}$ exposure and all-cause mortality, while difficult to determine from large-scale observational studies, is crucial for informing policy decisions. Various approaches have been developed to confront this challenge, and our analysis represents one of the first comprehensive comparisons of these approaches under diverse data-generating mechanisms that reflect real-world scenarios in air pollution epidemiology. Our analysis was designed to be accessible to the public, and the code for each step of the analysis is available on GitHub. Furthermore, we applied each method to data from the Medicare database, which represents the largest cohort to date encompassing the contiguous United States.

## 4.1 Simulation

**Evaluating regression estimators**

We briefly summarize our findings and contextualize them within the existing literature. In our simulations, we initially examined commonly employed regression methods, including multivariate linear regression, generalized additive models with smooth terms for nonlinear effects, and change point models for detecting threshold effects. Each of these regression methods does not explicitly separate a design and analysis stage and is not formulated in a causal inference framework (Imbens and Rubin, 2015). Regression methods tend to perform well under two conditions: when the relationship between exposure and outcome is consistent with the regression model and there are no heterogeneous treatment effects (i.e. the outcome model is correctly specified). For instance, when the true relationship between exposure and outcome is linear ($\mu_{\text{linear}}$) the linear estimator exhibits the lowest absolute bias and RMSE across all exposure scenarios. This finding is consistent with standard linear model theory, where the ordinary least squares estimator is considered the best linear unbiased estimator of the average treatment effect when the outcome model is specified correctly (Lehmann et al., 2006). However, our simulations explore various exposure models that introduce different levels of confounding. Notably, even in scenarios where the exposure is a nonlinear function of confounders or contains an interaction ($E_{\text{nonlinear}}, E_{\text{interaction}}$), the linear estimator remains unbiased in our analyses. Similarly, the change point regression method exhibited low absolute bias and RMSE when the ERC takes the form of a threshold function ($\mu_{\text{threshold}}$).

However, in our simulation, regression methods proved inadequate when heterogeneous treatment effects were present, particularly in cases involving nonlinear or interactive relationships between exposure and confounders (e.g., $E_{\text{nonlinear}}, E_{\text{interaction}}$). When heterogeneous treatment effects were present, applying naive regression models often led to the largest absolute bias. Given the relatively low variance in regression methods, especially linear models, we often see them performing with RMSE comparable to causal inference methods at low sample sizes, even in the presence of substantial bias. Although comparable RMSE might convey similar performance, an examination of the ERCs reveals that regression methods misrepresented the exposure-response relationship. This finding underscores the importance of caution when using single metrics to quantify ERCs. In the context of our simulation study, we demonstrate that regression methods are appropriate only if the functional form of the ERC is known and if there is no interaction between any of the covariates and the exposure in generating the outcome.

**Evaluating causal inferences estimators**

Our second class of estimators employed entropy balancing weights in a linear model, a generalized additive model and a change point model for the outcome. In our simulations, entropy weighting methods emerged as the optimal estimator in the presence of heterogeneous treatment effects when the exposure model takes the form of a linear function of the confounders ($E_{\text{linear}}$, $E_{\text{heavy tail}}$). Additionally, as the sample size in our simulation increased, the absolute bias and RMSE of entropy weighting estimators decreased. For example, in the linear ERC setting, we observed that the weighting methods performed well in the presence of heterogeneity ($\mu_{\text{linear,int}}$) under $E_{\text{linear}}$ and $E_{\text{heavy tail}}$, showing reduced bias compared to regression methods when the sample sizes exceeded 200. Larger sample sizes can allow better covariate balance with a lower likelihood of generating extreme weights (Tübbicke, 2022). Extreme weights can introduce greater variability in the estimated ERC (Lee et al., 2011). In the absence of heterogeneous effects, the entropy balancing methods performed comparably to their regression counterparts. It is worth noting that entropy balancing weights tend to yield larger RMSE in these settings, likely due to the fact that balancing weights are constructed prior to regression, reducing bias but introducing variability into estimates (Golinelli et al., 2012).

Nonetheless, entropy balancing methods did not produce unbiased results when the exposure followed a nonlinear function of confounders $E_{\text{nonlinear}}$, $E_{\text{interaction}}$. For example, in the linear ERC scenario $\mu_{\text{linear}}$, the bias and RMSE increased for the entropy balancing methods compared to the unweighted regression methods. While entropy balancing effectively eliminates the correlation between exposure and all covariates (Figure 2), this alone may not guarantee satisfactory covariate balance (Yiu et al., 2018). Removing the correlation between covariates and exposure eliminates the linear association between the covariate means and the continuous treatment, but it does not necessarily imply independence between the reweighted covariates and exposure. To mitigate this problem, Tübbicke recommends the additional condition of rendering higher moments of exposure uncorrelated with the covariates (Tübbicke, 2022). We refer to the method that requires the second moment of the exposure to be uncorrelated with the covariates as second-moment entropy balancing. In Supplementary Section S3 we illustrate that this approach corrects the bias noted in the $E_{\text{nonlinear}}$ and $E_{\text{interaction}}$ settings. This informs our use of second-moment entropy balancing in our data application. However, it is important to note that second-moment entropy balancing might not always converge at smaller sample sizes and, when not required, it can slightly elevate the RMSE as compared to first-order entropy balancing (see supplementary materials).

Finally, we evaluated the performance of the continuous matching causal inference method implemented in the CausalGPS package. Generally, the continuous matching estimator exhibited poor performance when the sample size was small. However, its performance improved significantly as the sample size increased, especially in the presence of heterogeneous treatment effects and complex exposure scenarios. Among the seven estimators, only CausalGPS

and second-moment entropy balancing performed well in the scenarios $E_{\text{nonlinear}}$ and $E_{\text{interaction}}$ with heterogeneous treatment effects. However, CausalGPS demonstrated greater variability in the ERC fit and exhibited higher RMSE values compared to other estimators. The superior performance of the CausalGPS method at larger sample sizes is not surprising, as it employs gradient boosting to specify the generalized propensity score (GPS). Gradient boosting allows for nonlinearities and interactions when estimating the relationship between confounders and the exposure, and a larger sample size provides more data for training the machine learning algorithm, resulting in improved GPS approximation. Moreover, larger sample sizes enhance the ability of the CausalGPS algorithm to identify appropriate matches for imputing potential outcomes.

**Bias persistence in sublinear ERC setting**

We observed unexpected results in the sublinear ERC setting $\mu_{\text{sublinear}}$. Surprisingly, the GAM entropy model exhibited high bias and RMSE under heterogeneous treatment effects in the $E_{\text{linear}}$ and $E_{\text{heavy tail}}$ setting. We anticipated that the entropy balancing weights would reduce the bias of the regression estimators, but the bias persisted. One possible explanation for this observation is the placement of knots for the splines. In our study, the nonlinear component of the sublinear ERC occurs at lower levels of exposure, while most GAM algorithms distribute knots at equally spaced quantiles of the data. Since we fixed the number of knots in our GAM estimator to four (refer to supplementary materials), it is likely that the knot placement is insufficient to capture the non-linearity that occurs at lower exposure levels. We may expect improved performance if we include more knots at the exposure values where we anticipate nonlinear behavior to occur.

**Data-informed simulation**

Our additional simulation using Medicare data corroborates the findings of our primary simulations. In scenarios without an interaction term, regression methods demonstrated effectiveness, contingent on the accurate specification of the exposure-outcome relationship. However, in cases with heterogeneous treatment effects, causal inference methods showed a reduction in bias. For the $\mu_{\text{linear,int}}$ and $\mu_{\text{threshold,int}}$ scenarios, employing entropy weights proved beneficial in mitigating bias. This effect was not observed in the $\mu_{\text{sublinear,int}}$ setting, aligning with our primary simulation results. The RMSE of the causal inference methods generally is in parallel with that of the regression methods, although with a slight increase likely attributable to the additional variability introduced by weight determination during the design phase. Although the relationship between exposure and the six covariates is not explicitly defined in the data-informed simulation, the results seem consistent with either the linear or heavy-tailed exposure scenarios observed in our primary simulations. The congruence observed in this additional simulation and our primary simulations lends further support to the applicability of the findings of our simulation study in empirical settings.

**Key insights from simulation**

Our findings underscore several important considerations for researchers interested in assessing the effect of a continuous exposure and generating an ERC. The first consideration pertains to the presence of heterogeneous treatment effects. We generally find that regression methods, when not incorporated within a causal inference framework, are unsuitable when anticipating heterogeneous treatment effects. In a regression setting, interactions are not accounted for in the model specification, whereas achieving covariate balance mitigates this dependency. However, if there is limited or no evidence of heterogeneous treatment effects, regression methods can be a viable framework, particularly at smaller sample sizes. As the sample size increases, we recommend adopting a causal inference framework, such as entropy balancing weights, to address potential confounding. Causal inference methods, such as weighting and continuous matching, separate the design and analysis stages, improving the objectivity of the outcome data analysis (Rubin, 2008).

When using entropy balancing weights, it is important to carefully consider rendering higher moments of the exposure or treatment uncorrelated with the covariates to ensure the absence of residual confounding. In our analysis, we initially only required uncorrelatedness between the first moment of exposure and the covariates. However, this method proved to be inadequate in the $E_{\text{nonlinear}}$ and $E_{\text{interaction}}$ scenarios. As demonstrated in the supplementary materials, requiring additional uncorrelatedness between the second moment of exposure and covariates (second-moment entropy balancing) ameliorated this issue, provided that the sample size was sufficiently large to achieve convergence. Based on these results and previous literature, we propose a two-step process: first, employing first-moment entropy weighting and then assessing remaining dependencies by estimating the ERC between the exposure and covariates using the generated weighting scheme. If the resulting ERC is completely flat and the derivative is zero, this indicates sufficient balance in the weighting scheme (Tübbicke, 2022). If this is not the case, we recommend incorporating higher moment entropy balancing by requiring that higher moments of the exposure be uncorrelated with the covariates. It is important to exercise caution when increasing the number of moments of the exposure required to be uncorrelated with covariates, as it can reduce bias but increase the variance of the estimator (Tübbicke, 2022). Generally, we recommend implementing second-moment entropy balancing as endorsed by Tübbicke (Tübbicke, 2022) given its efficacy in our simulations.

In the context of a large sample size where the functional form of the underlying ERC curve is unknown, we recommend employing the continuous matching framework provided by the CausalGPS estimator or the GAM entropy weighting scheme that requires uncorrelatedness between first- and second-moments of the exposure and the covariates. The CausalGPS estimator demonstrates robustness against misspecification of either the GPS or outcome models, and larger sample sizes enable machine learning methods to more accurately estimate the GPS. Moreover, the matching step reduces the dependence between the exposure and potential confounders, resulting in causal effect estimates that are less influenced by choices made during outcome modeling (Wu, Mealli, et al., 2022). However, it is important to note that we observed considerable variability and inadequate covariate balance when using the CausalGPS estimator with smaller sample sizes. As a result, we do not recommend using it in this setting.

## 4.2 Data application

In our data application, we conducted the first application of five statistical approaches to a data set containing more than 500 million person years of Medicare data, with the objective of estimating the impact of annual $PM_{2.5}$ concentration on all-cause mortality (Figures 6, 7). Our findings reveal several notable results. Each of the estimators demonstrated an increase in all-cause mortality in relation to annual average $PM_{2.5}$ concentration, consistent with previous research (Josey et al., 2023; Wu, Mealli, et al., 2022; Wu, Braun, et al., 2020). From our nonlinear estimators (GAM, GAM entropy, CausalGPS), we observed evidence of a nonlinear association between log mortality and $PM_{2.5}$, indicating the need to relax the linearity assumption. Furthermore, based on the results of our simulation study, the ERC generated by the GAM entropy and CausalGPS estimators seem most plausible, given the substantial sample size of our data, the presence of a nonlinear relationship between exposure and outcome, and the potential existence of heterogeneous treatment effects (Josey et al., 2023). Both the GAM entropy and CausalGPS estimated ERCs displayed a steep rise in the mortality rate at lower $PM_{2.5}$ concentrations, followed by a gradual plateau at higher concentrations. The GAM entropy and CausalGPS estimators exhibited a sublinear relationship between exposure and outcome at concentrations below 9 $\mu g/m^3$, though the ERC generated by the CausalGPS estimator leveled off at a lower concentration of annual average $PM_{2.5}$. Although the change point models failed to converge on this data set, the sublinear shape exhibited by the nonlinear causal inference methods at low concentrations suggests that a threshold model is unlikely.

Regarding the relative mortality rate associated with a decrease in annual average $PM_{2.5}$ in relation to the current NAAQS annual standard of 12 $\mu g/m^3$, the CausalGPS estimator projected a more moderate reduction in relative mortality for concentrations ranging from 9-12 $\mu g/m^3$ compared to other estimators (Figure 7). Once again, considering our simulations, the GAM entropy and CausalGPS estimators appeared most plausible, providing varying magnitudes of evidence to establish a lower NAAQS on annual average $PM_{2.5}$. The shape of the relative mortality curve for CausalGPS is consistent with other results (Josey et al., 2023) and offers evidence of a sublinear relationship between $PM_{2.5}$ and all-cause mortality. This analysis contributes further evidence to support the implementation of stringent standards for $PM_{2.5}$ emissions, suggesting that a lower standard could prevent a significant number of premature deaths.

## 4.3 Conclusion

There are several limitations to our analysis. The primary limitation is that the evaluation of different estimators of the ERC was conducted using simulations. Our study was restricted to examining a limited number of scenarios, and it is possible that the results may vary under different data-generating processes. However, we included various specifications of the ERC, covariates, and confounding scenarios that we believe represent a wide range of plausible situations. Another limitation is the choice of estimators. Numerous estimators are available in the literature on regression and causal inference. We selected our estimators based on their availability in standard software and their widespread usage in practice. In our study, we evaluated the impact of the annual average $PM_{2.5}$ on all-cause mortality in the Medicare population. Like all non-experimental data methods, we relied on the conditional independence assumption, which asserts that all factors affecting both the exposure and outcome have been considered, implying no unmeasured confounders. This assumption is crucial, but unverifiable. To address potential unmeasured confounders, we include year indicators for time-varying factors and census geographic region indicators for spatially varying factors. However, unmeasured confounding may still affect our conclusions. Prior studies utilizing the same data set (Wu, Braun, et al., 2020; F. Dominici et al., 2022) used the E-value method (VanderWeele et al., 2017) to gauge sensitivity to unmeasured confounding. The E-value indicates the minimum strength that an unmeasured confounder must have with both the exposure and the outcome to nullify the observed association. In these analyses, such a confounder would require at least a 1.32-fold risk ratio with long-term $PM_{2.5}$ exposure and mortality to counteract the estimated effects. Despite differences in methodology, these insights lend confidence to the robustness of our conclusions against unmeasured confounding.

Determining the causal relationship between a continuous exposure and outcome is a critical scientific endeavour. In the context of air pollution epidemiology, understanding the shape of the ERC between air pollution and adverse health effects has significant policy implications. Various modeling strategies have been employed to assess this relationship, and many researchers have advocated for the development and implementation of causal inference methods to inform air pollution policies (Goldman et al., 2019; Peters et al., 2019; Carone et al., 2020). Although limited work has compared the performance of these different estimators, our analysis provides valuable guidance for researchers in choosing an appropriate method to estimate these consequential exposure-response curves accurately.

# 5 Acknowledgements

This research was supported by grants from the National Institutes of Health (Grant No. T32 ES007142, R01ES026217, R01MD012769, R01ES028033, 1R01ES030616, 1R01AG066793, 1R01ES029950, 1R01ES 034373-01) and the Alfred P. Sloan Foundation (Grant No. G-2020-13946, 1R01AG066793). We thank these funding agencies for their support.

# References


Ai, Chunrong, Oliver Linton, and Zheng Zhang (2022). "Estimation and Inference for the Counterfactual Distribution and Quantile Functions in Continuous Treatment Models". In: *Journal of Econometrics*. Annals Issue: In Honor of Ron Gallant 228.1, pp. 39–61. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2020.12.009. URL: https://www.sciencedirect.com/science/article/pii/S0304407621000543.

Ameircan Lung Association (2022). *State of the Air 2022*. American Lung Association. URL: https://www.lung.org/research/sota/city-rankings.

Balzer, Laura B et al. (2019). "A New Approach to Hierarchical Data Analysis: Targeted Maximum Likelihood Estimation for the Causal Effect of a Cluster-Level Exposure". In: *Statistical Methods in Medical Research* 28.6, pp. 1761–1780. ISSN: 0962-2802. DOI: 10.1177/0962280218774936. URL: https://doi.org/10.1177/0962280218774936.

Bickel, P. J., F. Götze, and W. R. van Zwet (2012). "Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses". In: *Selected Works of Willem van Zwet*. Ed. by Sara van de Geer and Marten Wegkamp. Selected Works in Probability and Statistics. New York, NY: Springer, pp. 267–297. ISBN: 978-1-4614-1314-1. DOI: 10.1007/978-1-4614-1314-1_17. URL: https://doi.org/10.1007/978-1-4614-1314-1_17.

Bind, Marie-Abèle (2019). "Causal Modeling in Environmental Health". In: *Annual review of public health* 40, pp. 23–43. ISSN: 0163-7525. DOI: 10.1146/annurev-publhealth-040218-044048. pmid: 30633715. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6445691/.

Bind, Marie-Abele C and Donald B Rubin (2019). "Bridging Observational Studies and Randomized Experiments by Embedding the Former in the Latter". In: *Statistical Methods in Medical Research* 28.7, pp. 1958–1978. ISSN: 0962-2802. DOI: 10.1177/0962280217740609. URL: https://doi.org/10.1177/0962280217740609.

Brewer, Dylan, Daniel Dench, and Laura O. Taylor (2023). "Advances in Causal Inference at the Intersection of Air Pollution and Health Outcomes". In: *Annual Review of Resource Economics* 15.1, pp. 455–469. DOI: 10.1146/annurev-resource-101722-081026. URL: https://doi.org/10.1146/annurev-resource-101722-081026.

Burnett, Richard et al. (2018). "Global Estimates of Mortality Associated with Long-Term Exposure to Outdoor Fine Particulate Matter". In: *Proceedings of the National Academy of Sciences* 115.38, pp. 9592–9597. DOI: 10.1073/pnas.1803222115. URL: https://www.pnas.org/doi/10.1073/pnas.1803222115.

Callaway, Brantly and Weige Huang (2020). "Distributional Effects of a Continuous Treatment with an Application on Intergenerational Mobility". In: *Oxford Bulletin of Economics and Statistics* 82.4, pp. 808–842. ISSN: 1468-0084. DOI: 10.1111/obes.12355. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/obes.12355.

Carone, Marco, Francesca Dominici, and Lianne Sheppard (2020). "In Pursuit of Evidence in Air Pollution Epidemiology: The Role of Causally Driven Data Science". In: *Epidemiology (Cambridge, Mass.)* 31.1, pp. 1–6. ISSN: 1044-3983. DOI: 10.1097/EDE.0000000000001090. pmid: 31430263. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6889002/.

Chambers, John M. and Trevor J. Hastie (1992). *Statistical Models in S*. New York: Routledge.

Chartered Clean Air Scientific Advisory Committee (2018). *Summary Minutes of the U.S. EPA CASAC Public Teleconference on Particulate Matter*.

Colangelo, Kyle and Ying-Ying Lee (2022). *Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments*. DOI: 10.48550/arXiv.2004.03036. arXiv: 2004.03036 [econ]. URL: http://arxiv.org/abs/2004.03036. preprint.

Crump, Richard K. et al. (2009). "Dealing with Limited Overlap in Estimation of Average Treatment Effects". In: *Biometrika* 96.1, pp. 187–199. ISSN: 0006-3444. DOI: 10.1093/biomet/asn055. URL: https://doi.org/10.1093/biomet/asn055.

Davenport, Coral (2023). "Biden Administration Moves to Tighten Limits on Deadly Air Pollution". In: *The New York Times. Climate*. ISSN: 0362-4331. URL: https://www.nytimes.com/2023/01/06/climate/epa-soot-pollution-biden.html.

Di, Qian, Heresh Amini, et al. (2019). "An Ensemble-Based Model of PM2.5 Concentration across the Contiguous United States with High Spatiotemporal Resolution". In: *Environment International* 130, p. 104909. ISSN: 0160-4120. DOI: 10.1016/j.envint.2019.104909. URL: https://www.sciencedirect.com/science/article/pii/S0160412019300650.

Di, Qian, Lingzhen Dai, et al. (2017). "Association of Short-term Exposure to Air Pollution With Mortality in Older Adults". In: *JAMA* 318.24, pp. 2446–2456. ISSN: 0098-7484. DOI: 10.1001/jama.2017.17923. URL: https://doi.org/10.1001/jama.2017.17923.

Di, Qian, Yan Wang, et al. (2017). "Air Pollution and Mortality in the Medicare Population". In: *New England Journal of Medicine* 376.26, pp. 2513–2522. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1702747. pmid: 28657878. URL: https://doi.org/10.1056/NEJMoa1702747.

Dominici, F. et al. (2022). "Assessing Adverse Health Effects of Long-Term Exposure to Low Levels of Ambient Air Pollution: Implementation of Causal Inference Methods". In: *Research Report (Health Effects Institute)* 2022.211, pp. 1–56. ISSN: 1041-5505. pmid: 36193708.

Dominici, Francesca and Corwin Zigler (2017). "Best Practices for Gauging Evidence of Causality in Air Pollution Epidemiology". In: *American Journal of Epidemiology* 186.12, pp. 1303–1309. ISSN: 0002-9262. DOI: 10.1093/aje/kwx307. pmid: 29020141. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5860368/.

Dong, Nianbo et al. (2020). "Using Propensity Score Analysis of Survey Data to Estimate Population Average Treatment Effects: A Case Study Comparing Different Methods". In: *Evaluation Review* 44.1, pp. 84–108. ISSN: 0193-841X, 1552-3926. DOI: 10.1177/0193841X20938497. URL: http://journals.sagepub.com/doi/10.1177/0193841X20938497.

EPA (2019). *Integrated Science Assessment (ISA) for Particulate Matter*. EPA.

Fong, Christian, Chad Hazlett, and Kosuke Imai (2018). "Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements". In: *The Annals of Applied Statistics* 12.1, pp. 156–177. ISSN: 1932-6157. JSTOR: 26542524. URL: https://www.jstor.org/stable/26542524.

Fong, Youyi et al. (2017). "Chngpt: Threshold Regression Model Estimation and Inference". In: *BMC Bioinformatics* 18.1, p. 454. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1863-x. URL: https://doi.org/10.1186/s12859-017-1863-x.

Goldman, Gretchen T. and Francesca Dominici (2019). "Don't Abandon Evidence and Process on Air Pollution Policy". In: *Science* 363.6434, pp. 1398–1400. DOI: 10.1126/science.aaw9460. URL: https://www.science.org/doi/full/10.1126/science.aaw9460.

Golinelli, Daniela et al. (2012). "Bias and Variance Trade-Offs When Combining Propensity Score Weighting and Regression: With an Application to HIV Status and Homeless Men". In: *Health services & outcomes research methodology* 12.2-3, pp. 104–118. ISSN: 1387-3741. pmid: 22956891. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3433039/.

Greifer, Noah (2022). *WeightIt: Weighting for Covariate Balance in Observational Studies*. R package version 0.13.1. URL: https://CRAN.R-project.org/package=WeightIt.

Hainmueller, Jens (2012). "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies". In: *Political Analysis* 20.1, pp. 25–46. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/mpr025. URL: https://www.cambridge.org/core/journals/political-analysis/article/abs/entropy-balancing-for-causal-effects-a-multivariate-reweighting-method-to-produce-balanced-samples-in-observational-studies/220E4FC838066552B53128E647E4FAA7.

Health Effects Institute (2020). *State of Global Air 2020*. Special Report. Boston, MA:Health Effects Institute: Health Effects Institute. URL: https://www.stateofglobalair.org/.

Heckman, James J. and Richard Robb (1985). "Alternative Methods for Evaluating the Impact of Interventions: An Overview". In: *Journal of Econometrics* 30.1, pp. 239–267. ISSN: 0304-4076. DOI: 10.1016/0304-4076(85)90139-3. URL: https://www.sciencedirect.com/science/article/pii/0304407685901393.

Hernán, Miguel Ángel, Babette Brumback, and James M. Robins (2000). "Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men". In: *Epidemiology* 11.5, pp. 561–570. ISSN: 1044-3983. JSTOR: 3703998. URL: https://www.jstor.org/stable/3703998.

Hirano, Keisuke and Guido W. Imbens (2005). "The Propensity Score with Continuous Treatments". In: *Wiley Series in Probability and Statistics*. Ed. by Andrew Gelman and Xiao-Li Meng. Chichester, UK: John Wiley & Sons, Ltd, pp. 73–84. ISBN: 978-0-470-09045-9 978-0-470-09043-5. DOI: 10.1002/0470090456.ch7. URL: https://onlinelibrary.wiley.com/doi/10.1002/0470090456.ch7.

Huber, Martin et al. (2020). "Direct and Indirect Effects of Continuous Treatments Based on Generalized Propensity Score Weighting". In: *Journal of Applied Econometrics* 35.7, pp. 814–840. ISSN: 1099-1255. DOI: 10.1002/jae.2765. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2765.

Huling, Jared D., Noah Greifer, and Guanhua Chen (2023). "Independence Weights for Causal Inference with Continuous Treatments". In: *Journal of the American Statistical Association*, pp. 1–25. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2023.2213485. URL: https://www.tandfonline.com/doi/full/10.1080/01621459.2023.2213485.

Imai, Kosuke and David A van Dyk (2004). "Causal Inference With General Treatment Regimes". In: *Journal of the American Statistical Association* 99.467, pp. 854–866. ISSN: 0162-1459. DOI: 10.1198/016214504000001187. URL: https://doi.org/10.1198/016214504000001187.

Imbens, Guido W. (2004). "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review". In: *The Review of Economics and Statistics* 86.1, pp. 4–29. ISSN: 0034-6535. DOI: 10.1162/003465304323023651. URL: https://doi.org/10.1162/003465304323023651.

— (2000). "The Role of the Propensity Score in Estimating Dose-Response Functions". In: *Biometrika* 87.3, pp. 706–710. ISSN: 0006-3444. DOI: 10.1093/biomet/87.3.706. URL: https://doi.org/10.1093/biomet/87.3.706.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press. 647 pp. ISBN: 978-0-521-88588-1. Google Books: Bf1tBwAAQBAJ.

Josey, Kevin P. et al. (2023). "Air Pollution and Mortality at the Intersection of Race and Social Class". In: *New England Journal of Medicine* 388.15, pp. 1396–1404. ISSN: 0028-4793. DOI: 10.1056/NEJMsa2300523. URL: https://doi.org/10.1056/NEJMsa2300523.

Kennedy, Edward H. et al. (2017). "Non-Parametric Methods for Doubly Robust Estimation of Continuous Treatment Effects". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 79.4, pp. 1229–1245. ISSN: 1369-7412. JSTOR: 26773159. URL: https://www.jstor.org/stable/26773159.

Kreif, Noémi et al. (2015). "Evaluation of the Effect of a Continuous Treatment: A Machine Learning Approach with an Application to Treatment for Traumatic Brain Injury". In: *Health Economics* 24.9, pp. 1213–1228. ISSN: 1099-1050. DOI: 10.1002/hec.3189. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.3189.

Lechner, Michael (2001). "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption". In: *Econometric Evaluation of Labour Market Policies*. Ed. by Michael Lechner and Friedhelm Pfeiffer. ZEW Economic Studies. Heidelberg: Physica-Verlag HD, pp. 43–58. ISBN: 978-3-642-57615-7. DOI: 10.1007/978-3-642-57615-7_3.

Lechner, Michael and Anthony Strittmatter (2019). "Practical Procedures to Deal with Common Support Problems in Matching Estimation". In: *Econometric Reviews* 38.2, pp. 193–207. ISSN: 0747-4938. DOI: 10.1080/07474938.2017.1318509. URL: https://doi.org/10.1080/07474938.2017.1318509.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart (2011). "Weight Trimming and Propensity Score Weighting". In: *PLOS ONE* 6.3, e18174. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0018174. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018174.

Lehmann, Erich L. and George Casella (2006). *Theory of Point Estimation*. Springer Science & Business Media. 610 pp. ISBN: 978-0-387-22728-3. Google Books: 4f24CgAAQBAJ.

Liu, Cong et al. (2019). "Ambient Particulate Air Pollution and Daily Mortality in 652 Cities". In: *New England Journal of Medicine* 381.8, pp. 705–715. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1817364. pmid: 31433918. URL: https://doi.org/10.1056/NEJMoa1817364.

Manisalidis, Ioannis et al. (2020). "Environmental and Health Impacts of Air Pollution: A Review". In: *Frontiers in Public Health* 8. ISSN: 2296-2565. URL: https://www.frontiersin.org/articles/10.3389/fpubh.2020.00014.

Murray, Christopher J. L. et al. (2020). "Global Burden of 87 Risk Factors in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019". In: *The Lancet* 396.10258, pp. 1223–1249. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(20)30752-2. pmid: 33069327. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30752-2/fulltext.

Nasari, Masoud M. et al. (2016). "A Class of Non-Linear Exposure-Response Models Suitable for Health Impact Assessment Applicable to Large Cohort Studies of Ambient Air Pollution". In: *Air Quality, Atmosphere & Health* 9.8, pp. 961–972. ISSN: 1873-9326. DOI: 10.1007/s11869-016-0398-z. URL: https://doi.org/10.1007/s11869-016-0398-z.

National Academies of Sciences, Engineering, and Medicine (2022). *Advancing the Framework for Assessing Causality of Health and Welfare Effects to Inform National Ambient Air Quality Standard Reviews*. Washington, DC: The National Academies Press. URL: https://nap.nationalacademies.org/download/26612#.

Owens, Elizabeth Oesterling et al. (2017). "Framework for Assessing Causality of Air Pollution-Related Health Effects for Reviews of the National Ambient Air Quality Standards". In: *Regulatory Toxicology and Pharmacology* 88, pp. 332–337. ISSN: 0273-2300. DOI: 10.1016/j.yrtph.2017.05.014. URL: https://www.sciencedirect.com/science/article/pii/S0273230017301290.

Peters, Annette et al. (2019). "Promoting Clean Air: Combating Fake News and Denial". In: *The Lancet Respiratory Medicine* 7.8, pp. 650–652. ISSN: 2213-2600, 2213-2619. DOI: 10.1016/S2213-2600(19)30182-1. pmid: 31221566. URL: https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(19)30182-1/fulltext.

Petersen, Maya L et al. (2012). "Diagnosing and Responding to Violations in the Positivity Assumption". In: *Statistical methods in medical research* 21.1, pp. 31–54. ISSN: 0962-2802. DOI: 10.1177/0962280210386207. pmid: 21030422. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4107929/.

Politis, Dimitris N. and Joseph P. Romano (1994). "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions". In: *The Annals of Statistics* 22.4, pp. 2031–2050. ISSN: 0090-5364. JSTOR: 2242497. URL: https://www.jstor.org/stable/2242497.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Robins, James M. (2000). "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference". In: *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Ed. by M. Elizabeth Halloran and Donald Berry. The IMA Volumes in Mathematics and Its Applications. New York, NY: Springer, pp. 95–133. ISBN: 978-1-4612-1284-3. DOI: 10.1007/978-1-4612-1284-3_2.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao (1994). "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed". In: *Journal of the American Statistical Association* 89.427, pp. 846–866. ISSN: 0162-1459. DOI: 10.1080/01621459.1994.10476818. URL: https://doi.org/10.1080/01621459.1994.10476818.

Rosembaum, PAUL R. and DONALD B. RUBIN (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In: *Biometrika* 70.1, pp. 41–55. ISSN: 0006-3444. DOI: 10.1093/biomet/70.1.41. URL: https://doi.org/10.1093/biomet/70.1.41.

Rubin, Donald B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies". In: *Journal of Educational Psychology* 66.5, pp. 688–701. ISSN: 1939-2176. DOI: 10.1037/h0037350.

— (2008). "For Objective Causal Inference, Design Trumps Analysis". In: *The Annals of Applied Statistics* 2.3, pp. 808–840. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/08-AOAS187. URL: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/For-objective-causal-inference-design-trumps-analysis/10.1214/08-AOAS187.full.

— (1980). "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment". In: *Journal of the American Statistical Association* 75.371, pp. 591–593. ISSN: 0162-1459. DOI: 10.2307/2287653. JSTOR: 2287653. URL: https://www.jstor.org/stable/2287653.

Rubin, Donald B. and Neal Thomas (1996). "Matching Using Estimated Propensity Scores: Relating Theory to Practice". In: *Biometrics* 52.1, pp. 249–264. ISSN: 0006-341X. DOI: 10.2307/2533160. JSTOR: 2533160. URL: https://www.jstor.org/stable/2533160.

Schulz, Juliana and Erica E. M. Moodie (2021). "Doubly Robust Estimation of Optimal Dosing Strategies". In: *Journal of the American Statistical Association* 116.533, pp. 256–268. ISSN: 0162-1459. DOI: 10.1080/01621459.2020.1753521. URL: https://doi.org/10.1080/01621459.2020.1753521.

Shi, Liuhua et al. (2021). "A National Cohort Study (2000–2018) of Long-Term Air Pollution Exposure and Incident Dementia in Older Adults in the United States". In: *Nature Communications* 12.1 (1), p. 6754. ISSN: 2041-1723. DOI: 10.1038/s41467-021-27049-2. URL: https://www.nature.com/articles/s41467-021-27049-2.

Tübbicke, Stefan (2022). "Entropy Balancing for Continuous Treatments". In: *Journal of Econometric Methods* 11.1, pp. 71–89. ISSN: 2156-6674. DOI: 10.1515/jem-2021-0002. URL: https://www.degruyter.com/document/doi/10.1515/jem-2021-0002/html.

US Environmental Protection Agency (2015). *Clean Air Act Overview: Setting Emissions Standards for Major Sources of Toxic Air Pollutants*. Clean Air Act Overview. URL: https://www.epa.gov/clean-air-act-overview/setting-emissions-standards-major-sources-toxic-air-pollutants.

Van der Laan, Mark J. and Sherri Rose (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-1-4419-9781-4 978-1-4419-9782-1. DOI: 10.1007/978-1-4419-9782-1. URL: https://link.springer.com/10.1007/978-1-4419-9782-1.

VanderWeele, Tyler J. and Peng Ding (2017). "Sensitivity Analysis in Observational Research: Introducing the E-Value". In: *Annals of Internal Medicine* 167.4, pp. 268–274. ISSN: 0003-4819. DOI: 10.7326/M16-2607. URL: https://www.acpjournals.org/doi/10.7326/m16-2607.

Vegetabile, Brian G. et al. (2021). "Nonparametric Estimation of Population Average Dose-Response Curves Using Entropy Balancing Weights for Continuous Exposures". In: *Health Services and Outcomes Research Methodology* 21.1, pp. 69–110. ISSN: 1572-9400. DOI: 10.1007/s10742-020-00236-2. URL: https://doi.org/10.1007/s10742-020-00236-2.

Ward-Caviness, Cavin K. et al. (2021). "Long-Term Exposure to Particulate Air Pollution Is Associated With 30-Day Readmissions and Hospital Visits Among Patients With Heart Failure". In: *Journal of the American Heart Association* 10.10, e019430. DOI: 10.1161/JAHA.120.019430. URL: https://www.ahajournals.org/doi/full/10.1161/JAHA.120.019430.

Wei, Yaguang et al. (2020). "Causal Effects of Air Pollution on Mortality Rate in Massachusetts". In: *American Journal of Epidemiology* 189.11, pp. 1316–1323. ISSN: 0002-9262. DOI: 10.1093/aje/kwaa098. URL: https://doi.org/10.1093/aje/kwaa098.

Wood, Simon N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press. 497 pp. ISBN: 978-1-4987-2834-8.

World Health Organization (2021). *WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. World Health Organization. xxi, 273. ISBN: 978-92-4-003422-8. URL: https://apps.who.int/iris/handle/10665/345329.

Wu, Xiao, Danielle Braun, et al. (2020). "Evaluating the Impact of Long-Term Exposure to Fine Particulate Matter on Mortality among the Elderly". In: *Science Advances* 6.29, eaba5692. DOI: 10.1126/sciadv.aba5692. URL: https://www.science.org/doi/full/10.1126/sciadv.aba5692.

Wu, Xiao, Fabrizia Mealli, et al. (2022). "Matching on Generalized Propensity Scores with Continuous Exposures". In: *Journal of the American Statistical Association* 0.0, pp. 1–29. ISSN: 0162-1459. DOI: 10.1080/01621459.2022.2144737. URL: https://doi.org/10.1080/01621459.2022.2144737.

Yiu, Sean and Li Su (2018). "Covariate Association Eliminating Weights: A Unified Weighting Framework for Causal Effect Estimation". In: *Biometrika* 105.3, pp. 709–722. ISSN: 0006-3444. DOI: 10.1093/biomet/asy015. URL: https://doi.org/10.1093/biomet/asy015.

Zanutto, Elaine (2006). "A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data". In: *Journal of Data Science* 4, pp. 67–91.

Zhao, Shandong, David A van Dyk, and Kosuke Imai (2020). "Propensity Score-Based Methods for Causal Inference in Observational Studies with Non-Binary Treatments". In: *Statistical Methods in Medical Research* 29.3, pp. 709–727. ISSN: 0962-2802. DOI: 10.1177/0962280219888745. URL: https://doi.org/10.1177/0962280219888745.

Zhu, Yeying, Donna L. Coffman, and Debashis Ghosh (2015). "A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments". In: *Journal of Causal Inference* 3.1, pp. 25–40. ISSN: 2193-3685. DOI: 10.1515/jci-2014-0022. URL: https://www.degruyter.com/document/doi/10.1515/jci-2014-0022/html?lang=de.

Zigler, Corwin Matthew et al. (2016). "Causal Inference Methods for Estimating Long-Term Health Effects of Air Quality Regulations". In: *Research Report (Health Effects Institute)* 187, pp. 5–49. ISSN: 1041-5505. pmid: 27526497.