Estimating a Causal Exposure Response Function with a Continuous Error-Prone Exposure: A Study of Fine Particulate Matter and All-Cause Mortality

Kevin P. Josey, Priyanka deSouza, Xiao Wu, Danielle Braun, and Rachel Nethery

Numerous studies have examined the associations between long-term exposure to fine particulate matter (PM$_{2.5}$) and adverse health outcomes. Recently, many of these studies have begun to employ high-resolution predicted PM$_{2.5}$ concentrations, which are subject to measurement error. Previous approaches for exposure measurement error correction have either been applied in non-causal settings or have only considered a categorical exposure. Moreover, most procedures have failed to account for uncertainty induced by error correction when fitting an exposure response function (ERF). To remedy these deficiencies, we develop a multiple imputation framework that combines regression calibration and Bayesian techniques to estimate a causal ERF. We demonstrate how the output of the measurement error correction steps can be seamlessly integrated into a Bayesian additive regression trees (BART) estimator of the causal ERF. We also demonstrate how kernel-weighted smoothing of the posterior samples from BART can be used to create a more accurate ERF estimate. Our proposed approach also properly propagates the exposure measurement error uncertainty to yield accurate standard error estimates. We assess the robustness of our proposed approach in an extensive simulation study. We then apply our methodology to estimate the effects of PM$_{2.5}$ on all-cause mortality among Medicare enrollees in New England from 2000 to 2012.

Key Words: Measurement Error; Causal Inference; Multiple Imputation; Air Pollution; Environmental Epidemiology.

Danielle Braun and Rachel Nethery are the co-senior authors for this article.

K. P. Josey (✉) · D. Braun · R. Nethery, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA (E-mail: kjosey@hsph.harvard.edu).

P. de Souza, Department of Urban and Regional Planning, University of Colorado, Denver, CO, USA.

X. Wu, Department of Statistics, Stanford University, Stanford, CA, USA; Stanford Data Science, Stanford University, Stanford, CA, USA.

D. Braun, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA.

© 2022 International Biometric Society, corrected publication 2023

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 1, Pages 20–41
https://doi.org/10.1007/s13253-022-00508-z

20
1. INTRODUCTION

Methods for conducting causal inference with continuous exposures have gained significant traction in recent years. The goal of these methods is to estimate a causal exposure response function (ERF) that is free of confounding bias (Kennedy et al. 2017; Wu et al. 2018). Estimators of an ERF typically rely on estimates of either the generalized propensity score (GPS), a marginalized model of the outcome process, or some combination of the two to adjust for confounding. However, most of these methods make the implicit assumption that the exposure is measured without error, which is implausible in many observational study settings. For example, in modern environmental health research, air pollution exposure measurements are often derived from model-based predictions of pollutant concentrations rather than the exact pollutant concentrations experienced by an individual. Moreover, because many studies record only areal measures of residential locations such as ZIP codes, cities, or counties, exposure measures often represent aggregated predictions across these areas. Using aggregated exposures assumes that concentrations are homogeneous within areas and experienced by all individuals residing in the area. As such, exposure measurement error is prevalent in many air pollution epidemiological studies (Kioumourtzoglou et al. 2014). Using an error-prone exposure (EPE) in place of the true exposure violates standard causal inference assumptions and may result in biased ERF estimates. Due to the policy relevance of air pollution ERFs and particularly those produced using causal inference techniques, invalid inferences caused by EPEs could have severe consequences.

While measurement error has been studied extensively outside of causal inference settings (Carroll et al. 2006; Cole et al. 2006), accounting for EPEs in causal inference is a relatively new endeavor and is mainly confined to scenarios with binary and categorical exposures (Lewbel 2007; Braun et al. 2017; Wu et al. 2019), or cases of measurement error in a confounder instead of the exposure (Lenis et al. 2017; Webb-Vargas et al. 2017). Beyond the issues encountered when using an EPE in a typical generalized linear model setting, accommodating an EPE in a causal model presents additional challenges in conjunction with resolving confounding bias (Braun et al. 2017), many of which remain unaddressed. Moreover, propagating measurement error-related uncertainty into ERF estimates is paramount for proper inference in not just causal settings, but more broadly for many environmental epidemiological studies.

We develop a multiple imputation framework for estimating causal ERFs, which incorporates corrections for various types of exposure measurement error commonplace in environmental epidemiology. This approach jointly samples from three sub-models: (1) a model of the EPE, which imputes the true exposures using information from covariates and validation data when available, (2) a GPS model, which improves the accuracy of the imputations, and (3) an outcome model, which estimates the ERF using the imputed exposures while adjusting for confounders. Markov chain Monte Carlo (MCMC) methods are used to sample from the posterior distributions of the unobserved true exposures and the expected outcome conditioned on the exposures and confounders. A Bayesian additive regression trees (BART) model is specified for the outcome sub-model to capture complex, nonlinear relationships between the exposure, the covariates, and the outcome. We also show that implementing a further smoothing step on the BART-estimated ERF using local regression techniques pro-
vides accurate estimates of the ERF with fewer posterior samples than a classic Bayesian approach, all while adequately propagating the measurement error uncertainty. The latter smoothing step transitions our approach from a Bayesian estimator into one better characterized as a multiple imputation estimator, while at the same time producing smoother estimates of the ERF than the notoriously noisy BART output (Nethery et al. 2021).

The remainder of this article is structured as follows: We describe the motivating data example in Sect. 1.1. In Sect. 2, we introduce notation, define the measurement error sources, and identify the causal ERF model. Section 3 outlines the procedure for correcting the attenuation bias created by measurement error using multiple imputation methods, and we discuss several important caveats and limitations to our proposed methodology. Section 4 contains a simulation study. Section 5 applies the proposed method to create a measurement error-corrected causal ERF for long-term fine particulate matter (PM$_{2.5}$) exposure and all-cause mortality in the Medicare population in New England, using the data described in Sect. 1.1. We conclude with a discussion in Sect. 6.

1.1. Motivating Example

PM$_{2.5}$ is a well-studied air pollutant known to adversely impact numerous health outcomes (Hajat et al. 2002; Dominici et al. 2006; Brook et al. 2010; Zhu et al. 2017), including all-cause mortality (Anderson 2009; Di et al. 2017; Wu et al. 2020), cardiovascular disease (Dominici et al. 2006; Zanobetti et al. 2009; Pope et al. 2015; Yazdi et al. 2019; Danesh Yazdi et al. 2021), and pulmonary/respiratory diseases (Dominici et al. 2006; Zanobetti et al. 2009; Rhee et al. 2019). In one of these studies, using a cohort of Medicare enrollees 2000–2012, Wu et al. (2019) implemented various causal inference techniques to assess whether long-term PM$_{2.5}$ exposure increases the risk of all-cause mortality among older Americans (> 65 years of age). While Wu et al. (2019) employed a regression calibration approach to resolve parts of the exposure measurement error present in this application, there were other error components that were left uncorrected. Moreover, a simple regression calibration approach may fail to adequately propagate measurement error-related uncertainty into the causal effect estimates. Additionally, Wu et al. (2019) considered exposure categories instead of examining the ERF across a continuum of exposures. Categorization of exposures implicitly assumes that participants in the same category are exposed to the same exposure level, which may induce yet another source of measurement error. We seek to investigate the same scientific question as Wu et al. (2019) using the same data but employing a novel analytic approach that: (1) corrects for additional sources of measurement error in the PM$_{2.5}$ exposures and (2) models the continuous ERF as opposed to categorizing the exposure.

We utilize annual average PM$_{2.5}$ predictions for 1km×1km grid-years across the USA from Di et al. (2016), which were generated via a neural network that combines information from satellite, ground monitor, and land-use data. These exposure predictions have been widely used in epidemiological studies (Di et al. 2017; Rhee et al. 2019; Wu et al. 2019; Yazdi et al. 2019; Wu et al. 2020; Danesh Yazdi et al. 2021). Model assessments indicate that the accuracy of the predictions varies across regions (Di et al. 2016). Because these exposure predictions are error-prone, any inference drawn from them without correction may be biased. To control for this source of measurement error, we use the same ground-monitor
Estimating a Causal Exposure Response

PM$_{2.5}$ data that were used to generate the grid-predicted PM$_{2.5}$ measurements in the first place. We treat the measurements from these monitors as the gold standard measurements of PM$_{2.5}$ exposure (i.e., error-free measurements). Our model uses the monitored data as a validation set to re-calibrate the error-prone PM$_{2.5}$ measurements. PM$_{2.5}$ ground monitor locations in New England are overlaid on a ZIP-code map in Fig. 1.

Another limitation of the data is that the residential locations of the Medicare enrollees can only be mapped to ZIP codes (Wu et al. 2020). Thus, following common practice in this setting, we use ZIP-code years as our units of analysis and all-cause mortality rates within ZIP-code years as our outcome. This creates a misalignment between the PM$_{2.5}$ exposure predictions, which occur on 1-km grids, and the ZIP-codes to which we wish to assign PM$_{2.5}$ exposures.

We also leverage data on demographic and socioeconomic status (SES) factors, as well as land-use variables which are used to aid in measurement error correction. The following potential confounder variables are collected from the US Census Bureau at the ZIP-code-
year level: population density, median household income, percent population and households below poverty level, racial/ethnic composition, and distribution of educational attainment (college, some college, high school, not completed high school). Additionally, individual-level information about the Medicare enrollees, including their sex, age, race, and Medicaid eligibility status (a proxy for low-income), are obtained from the Medicare record and summarized at the ZIP-code-year level. The land-use variables, collected at the grid-year level, are: surface temperature, accumulated precipitation, radiation flux, accumulated total evaporation, heat flux, precipitation rate, humidity, snow cover, cloud cover, and wind speed.

2. PRELIMINARIES

2.1. NOTATION AND MEASUREMENT ERROR

Our data involve measurements at two different spatial scales: ZIP-code-years (outcome, covariates) and 1km x 1km grid-years (exposure, land-use variables). In general, we refer to grid-years as cells, and cells are assumed to be nested within ZIP-code-years, which we refer to as clusters. Let $X_i \in \mathcal{X}$ be a vector of covariates for cluster $i$ ($i = 1, 2, \ldots, n$) and $Y_i$ denote the outcome variable in that cluster. While the methodology we present easily generalizes to other outcome distributions, in our applied example the $Y_i$ are counts (e.g., number of deaths in ZIP-code-year $i$) with $N_i$ representing an offset (e.g., number of at-risk person-years in ZIP-code-year $i$). The combination of these two measures allows us to define the rate $\bar{Y}_i \equiv Y_i / N_i$. We denote the cumulative density functions for the $X_i$ with $P(x)$ and the associated empirical density function with $\hat{P}(x)$.

We assume there is a true exposure both on the cluster level and on the cell level. To enable alignment with the outcome data at the cluster level, the true cluster-level exposure, denoted $A_i \in \mathcal{A}$, is the latent variable we are most concerned with accurately predicting. We assume that the $A_i$ are continuous but unobserved for all clusters. The true cell-level exposures within cluster $i$ are indexed by $j = 1, 2, \ldots, M_i$ and are denoted as $S_{ij}$. The total number of cells is written as $m \equiv \sum_{i=1}^{n} M_i$. We assume, as in our example, that the $S_{ij}$ will only be observed for a subset of cells—the grids containing monitors—which are indexed by $(i, j) \in S$. We also have for every cluster and cell an observed model-predicted value of $S_{ij}$, denoted as $\tilde{S}_{ij}$. These predictions are the error-prone exposures. The aggregate value of these EPEs to the cluster level will be denoted with $\bar{Z}_i = M_i^{-1} \sum_{j=1}^{M_i} \tilde{S}_{ij}$. Figure 2 illustrates the relationship between $A_i$, $S_{ij}$, and $\tilde{S}_{ij}$ in the context of our motivating application. We also allow for the existence of a set of cell-level features, denoted by the vector $W_{ij}$, that might inform us about the relationship between $S_{ij}$ and $\tilde{S}_{ij}$. In our applied example, the $W_{ij}$ are land-use variables measured at the grid-year level. $X_i$ may contain aggregated and/or transformed values of the variables contained in $W_{ij}$.

Finally, the conditional expected value of any given variable $D$ is denoted by $\mu_D(\cdot) \equiv \mathbb{E}(D|\cdot)$. The conditional probability density function of $D$ is denoted with $p(D|\cdot)$. The generalized propensity score (GPS) is the following particular conditional probability density function: $p(A_i|X_i)$. 

2.2. Measurement Error Model

We formulate the measurement error in this context as a special case of nondifferential classical error. As we noted in Sect. 2.1, we must contend with the problem that $S_{ij}$ is rarely observed, and in most areas only the EPE measurements, $\tilde{S}_{ij}$, are available. Moreover, conducting a cluster-level analysis relating $A_i$ to $\tilde{Y}_i$ requires measurements of $A_i$, not $S_{ij}$. Therefore, we must consider approaches for summarizing $S_{ij}$ while accounting for the possible confounding influence of $X_i$ and the varying sizes of $M_i$, and ensuring the uncertainty from this process is propagated into the final estimator.

Throughout this paper, we will assume that $\mathbb{E}(S_{ij}|A_i) = A_i$, which can be framed as treating the true cell level exposures $S_{ij}$ as replicate measures of the corresponding cluster-level exposure $A_i$, each measured with some amount of error. This ‘replicate measures’ conceptualization is commonly used in the presentation of classical measurement error methods. An additional dimension of complexity is added by the need to rely on cell-level exposure predictions, $\tilde{S}_{ij}$. This leads us to formulate the measurement error structure as a variant of the typical nondifferential classical measurement error, which can be decomposed into two components as follows:

$$U_{ij} = \underbrace{S_{ij} - \tilde{S}_{ij}}_{\text{Prediction}} + \underbrace{\tilde{S}_{ij} - A_i}_{\text{Aggregation}} = \underbrace{S_{ij} - A_i}_{\text{Classical}}.$$ (1)
Following conventions in the classical measurement error literature, we make the assumption that the measurement error \( U_{ij} \) is homoscedastic and conditionally independent such that:

**Assumption 1.** (Conditionally Independent Measurement Error) We assume \((U_{ij} \perp \perp U_{ij'} | A_i, W_{ij})\) for two cells \( j \neq j' \) in a given cluster \( i = 1, 2, \ldots, n \);

**Assumption 2.** (Homoscedastic Measurement Error) For all \( j = 1, 2, \ldots, M_i \) contained within every \( i = 1, 2, \ldots, n \), we assume the measurement error conditional on \( A_i \) has constant variance; \( \mathbb{V}(U_{ij} | A_i) \equiv \omega^2 \).

The prediction error component in (1) can be characterized as a Berkson error term (Haber et al. 2020). The aggregation error, on the other hand, is a type of classical measurement error, implying the composite measurement error term \( U_{ij} \) is also a type of classical error. To help cement this understanding, ignore for a moment the prediction error term in \( U_{ij} \). The goal of regression calibration is to replace classical measurement error with Berkson error. It has been noted in the regression calibration literature that Berkson error resulting from regression calibration often yields less bias than its classical error counterpart (Carroll et al. 2006). Our method, presented in the following sections, is developed in the same spirit, replacing \( U_{ij} \) with a Berkson error term via multiple imputation. Our method also allows the user to replace the prediction error component of \( U_{ij} \), if desired, but doing so essentially substitutes the original Berkson prediction error with another Berkson error term. That said, re-calibrating the predictions \( \tilde{S}_{ij} \) can still be useful if we suspect \( E[\tilde{S}_{ij} | A_i] \neq A_i \). In this latter case, if we assume \( E[S_{ij} | A_i] = A_i \) for all \((i, j) \in S\), then we can use validation data to find an unbiased predictor for \( S_{ij} \), denoted by \( \hat{S}_{ij} \equiv \hat{\mu}_S(S_{ij}, W_{ij}) \), which implies \( E[\hat{S}_{ij} | A_i] = A_i \).

As we will show in Sect. 3, corrections for both the classical measurement error between \( S_{ij} \) and \( A_i \) and any potential bias found in the predictions \( \tilde{S}_{ij} \) can be accommodated under the proposed multiple imputation framework. The imputations of the true exposures are drawn from an MCMC sampler using the full conditional likelihood in (2).

### 2.3. Assumptions and Identification

The methods we employ operate within the Neyman–Rubin causal model modified for continuous exposures Rubin (1974). Let \( \bar{Y}_i(a) \) be the outcome that would occur in cluster \( i \) if, possibly contrary to what is observed, it had received exposure level \( A_i = a \). We refer to the \( \bar{Y}_i(a) \) for any \( a \in A \) as the potential outcomes. For some exposure level \( a \) the objective analysis is to estimate the marginal ERF, \( \theta(a) \equiv E[\bar{Y}_i(a)] \). Unlike in the binary or categorical exposure setting, where there is only a finite number of unrealized potential outcomes, with a continuous exposure there is an infinite number of unrealized potential outcomes for every unit in the study. Despite this perceived challenge, it is relatively straightforward to translate the assumptions intended for a categorical exposure under the Neyman–Rubin causal model to a continuous exposure setting necessary to identify and estimate \( \theta(a) \).

To start, we invoke the stable unit treatment value assumption which consists of two conditions: (1) consistency and (2) no interference. Consistency refers to the notion that
\( \bar{Y}_i = \bar{Y}_i(A_i) \), i.e., each unit’s observed outcome is equal to its potential outcome evaluated at the true exposure experienced. We are careful to define this assumption in the setup to our problem, recognizing that \( A_i \) is unobserved. While unobserved, the true exposure \( A_i \) still exists, so consistency should hold. On the other hand, if we naively substitute \( \tilde{Z}_i \) for \( A_i \), then consistency is likely violated with respect to the EPE unless \( \tilde{Z}_i = A_i \) implying \( \bar{Y}_i(\tilde{Z}_i) = \bar{Y}_i(A_i) \) for all \( i \). As this assumption is unlikely to hold, it is safe to assume consistency is violated if measurement error-agnostic approaches are taken. No interference refers to the assumption that the exposure of one unit does not affect the potential outcomes of another unit. This condition is difficult to uphold in many environmental epidemiology studies, including our own applied example. Major air pollution emission sources can affect broad areas (e.g., many ZIP-codes), which may lead to instances where the exposure assignment of one ZIP code affects the outcome of another (Papadogeorgou et al. 2019). We acknowledge this limitation, yet addressing it is outside the scope of this work.

The fundamental problem in causal inference with any type of exposure is that \( \theta(a) \) is not intrinsically estimable since \( \bar{Y}_i(a) \) is not observed for every \( a \in A \). However, \( \theta(a) \) can still be estimated under the strongly ignorable exposure assumption. With the Neyman–Rubin causal model, this requires the following two assumptions (Kennedy et al. 2017):

**Assumption 3.** (Overlap/positivity) \( p(A_i|X_i) > p_{\text{min}} > 0 \).

**Assumption 4.** (Weak unconfoundedness) \( [\bar{Y}_i(a) \perp \perp A_i] | X_i \) for all \( a \in A \).

The overlap assumption requires the exposure assignment mechanism to be non-deterministic when conditioned on the covariates. In other words, the probability that a unit is exposed to any level of the exposure along the support is always greater than zero (or \( p_{\text{min}} \)). Weak unconfoundedness states that the potential outcome evaluated at \( a \) does not depend on the true exposure \( A_i \) when adjusted by the set of observed confounding variables. This implies there can be no unmeasured confounding. If all confounders for the relationship between \( \bar{Y}_i \) and \( A_i \) are measured, the latter assumption should hold even when \( A_i \) exists but is not observed. When these assumptions hold, it is valid to draw causal conclusions either with experimental or observational data and a continuous exposure (Gill and Robins 2001), and assuming \( A_i \) is known, the true causal ERF can be identified as

\[
\mathbb{E} \left\{ \mathbb{E} \left[ \bar{Y}_i | A_i = a, X_i \right] \right\} = \mathbb{E} \left\{ \mathbb{E} \left[ \bar{Y}_i(a) | X_i \right] \right\} = \mathbb{E} \left[ \bar{Y}_i(a) \right] = \theta(a)
\]

### 3. A MULTIPLE IMPUTATION APPROACH

#### 3.1. SAMPLING THE EXPOSURE AND NUISANCE PARAMETERS

We introduce our method by first specifying the joint likelihood function for \( \bar{Y}_i, A_i, \) and \( S_{ij} \), which incorporates the components needed to address the measurement error described
by (1). Supposing that every piece of data were observed, we have:

\[
\prod_{i=1}^{n} \left\{ p_{\tilde{Y}_i} (\tilde{Y}_i | A_i, X_i) \times p_A (A_i | X_i, \phi_i) \times p_\phi (\phi_i | V) \times \prod_{j=1}^{M_i} p_S (S_{ij} | A_i) \right\} .
\]  

(2)

From a modeling perspective, we assume that the latent exposure variables are conditionally independent and approximately normal in distribution with \(A_i | X_i, \phi_i \sim N(\mu_A (X_i, \phi_i), \sigma^2)\) and \(S_{ij} | A_i \sim N(A_i, \omega^2)\), in accordance with Assumptions 1 and 2. A random effect \(\phi_i\) is included in the model for \(A_i\) to control for spatial autocorrelation between the cluster-level exposures (Lee 2013). We assume \(\phi \equiv (\phi_1, \phi_2, \ldots, \phi_n)^T \sim N(0, v^2 Q^{-1}(V; \rho))\) where \(Q(V; \rho)\) is a precision matrix controlling the auto-correlation structure as proposed by Leroux et al. (2000) given a binary adjacency matrix \(V\) and a hyperparameter \(\rho\) controlling the correlation between adjacent clusters. If we assume the \(A_i\) are independent, then we can set \(\phi_i = 0\) and \(p_\phi (\phi_i | V) = 1\) for all \(i\). Because \(S_{ij}\) is missing, we can substitute \(\tilde{S}_{ij}\) for \(S_{ij}\) if we suppose \(\mathbb{E}[\tilde{S}_{ij} | A_i] = A_i\). However, when given a set of validation data (i.e., a subset where \(S_{ij}\) is observed), a more conservative approach is to instead assume \(\mathbb{E}[S_{ij} | A_i] = A_i\) and draw a sequence of posterior predictions for \(S_{ij}\) for all \(j = 1, 2, \ldots, M_i\) and \(i = 1, 2, \ldots, n\). To do this, we can append the model \(\prod_{(i,j) \in S} p_S (S_{ij}, \tilde{W}_{ij})\) to (2) and, supposing that \(S_{ij} | \tilde{S}_{ij}, \tilde{W}_{ij} \sim N (\mu_S (\tilde{S}_{ij}, \tilde{W}_{ij}), \tau^2)\), sample values of \(S_{ij}\). Note that this addition overspecifies the full-conditional likelihood in (2), so doing so is only possible when validation data are available.

We use the above likelihood throughout this manuscript to describe how to sample posterior draws of \(A_i\) contemporaneously with posterior predictions for \(\mu_{\tilde{Y}} (A_i, X_i)\). We will show in the next section how these posterior samples can be smoothed to create an estimate of the ERF. Relative to the standard regression calibration approach where only a single imputation of \(A_i\) is used in fitting the outcome model, using a set of posterior samples (i.e., multiple imputations) of the error-corrected exposures should better propagate uncertainty attributable to the exposure measurement error into the outcome model.

We use a fully Bayesian joint modeling approach for the measurement error, GPS, and outcome models. Parameter values are sampled from the posterior distribution either by Gibbs or Metropolis–Hastings sampling steps, with an added intermediate step to generate posterior predictive samples of the latent variables \(A_i\) and, if necessary, \(S_{ij}\). We refer to this intermediate step as the imputation stage, whereas the steps involving sampling the parameters are referred to as the analysis stage. See the full details of the sampling algorithm in Supplemental Section S1.

We must carefully consider the form of \(\mu_A (X_i, \phi_i)\) and \(\mu_S (\tilde{S}_{ij}, \tilde{W}_{ij})\) to properly address bias in the EPE model while also accounting for confounding. In our simulation study and data analysis in Sects. 4 and 5, we assume linear forms for \(\mu_A (X_i, \phi_i) = \phi_i + X_i^T \beta\) and \(\mu_S (\tilde{S}_{ij}, \tilde{W}_{ij}) = (\tilde{S}_{ij}, \tilde{W}_{ij}^T) \alpha\). To better capture nonlinear associations and better avoid issues with model misspecification, splines or Gaussian processes could alternatively be specified (Antonelli et al. 2020; Ren et al. 2021).

If we assume that \(\tilde{Y}_i\) is a linear function of \(A_i\) and \(X_i\), then the parameters determining \(p_{\tilde{Y}} (\tilde{Y}_i | A_i, X_i)\) can be drawn using a Gibbs sampling step, assuming the \(N_i\) are fixed and
known. However, we prefer to use a data-driven model for $p_{\tilde{Y}}(\tilde{Y}_i|A_i, X_i)$ and therefore $\mu_{\tilde{Y}}(A_i, X_i)$, because this model is perhaps the most essential component in (2) for finding accurate estimates of $\theta(a)$. To this end, we employ an iteratively updated BART model (Chipman et al. 2010). Letting $t = 1, 2, \ldots, T$ index the trees for a given iteration of the MCMC, the BART model assumes

$$\tilde{Y}_i = \mu_{\tilde{Y}}(A_i, X_i) + \epsilon_i \approx \sum_{t=1}^{T} g(A_i, X_i; F_t, G_t) + \epsilon_i. \quad (3)$$

Here, $g(\cdot)$ is a function that bins observations into groups with binary trees formed by the rules contained in $F_t$, and node means characterized by the set $G_t$. A BART model differs from other regression tree ensembling methods because of the priors placed on $\mu_{\tilde{Y}}(A_i, X_i)$. Prior samples are denoted by superscript $(k)$, $k = 1, 2, \ldots, K$, e.g., $\mu_{\tilde{Y}}(A_i^{(k)}, X_i) \approx \sum_{t=1}^{T_k} g(A_i^{(k)}, X_i; F_t^{(k)}, G_t^{(k)})$. In each iteration of our MCMC sampler, the posterior samples of the BART parameters are drawn conditional on the current posterior sample of the exposures $A_i^{(k)}$. The error term is assigned the distribution $\epsilon_i \sim \mathcal{N}(0, \psi^2 / N_i)$.

For model fitting, both in the simulation study in Sect. 4 and the illustrative example in Sect. 5, we assign conjugate inverse-gamma prior distributions to the variance parameters, $\omega^2, \tau^2, \sigma^2, \nu^2, \psi^2 \sim IG(0.001, 0.001)$, and set $\rho \sim U(0, 1)$. The regression coefficients for $\mu_S(\cdot), \mu_A(\cdot)$, are each assigned independent Gaussian priors with default values $\alpha, \beta \sim \mathcal{N}(0, 10^6 I)$ where $\theta$ is a vector with every entry equal to zero and $I$ is an identity matrix (with appropriate dimensions). Additional sampling details are provided in Supplemental Section S1. For an example of how to implement a generalized linear model for $\tilde{Y}_i|A_i^{(k)}, X_i$ instead of the BART implementation in (3), please refer to the simulation experiment in Supplemental Sect. S4.

### 3.2. Smoothing BART Output

While predictions from the BART model (e.g., means of posterior predictive samples) could be used directly to form an estimate of the ERF, these samples do not typically form a smooth function of $\theta(a)$ over $a \in \mathcal{A}$. Smooth ERFs are believed to be most biologically plausible relationship in many epidemiological and biomedical applications, including the effect of PM$_{2.5}$ on all-cause mortality, and are more desirable for identifying causal relationships (Kim et al. 2018). To resolve this problem, we will project a multiply-imputed pseudo-outcome derived from the BART output onto the support of $\mathcal{A}$ with local regression techniques. We only need a small number of imputations, i.e., a small, thinned subset of the BART posterior samples (indexed by $l = 1, 2, \ldots, L, L \ll K$), to get smooth, reliable estimates of the ERF. First, we create multiple imputations of the pseudo-outcome as

$$\xi^{(l)}(A_i^{(l)}, X_i, \tilde{Y}_i) = \left[ \tilde{Y}_i - \mu_{\tilde{Y}}^{(l)}(A_i^{(l)}, X_i) \right] + \int_{\mathcal{X}} \mu_{\tilde{Y}}^{(l)}(A_i^{(l)}, x) d\mathbb{P}(x). \quad (4)$$

There are two components in this pseudo-outcome. The integral on the right-hand side of the addition symbol is the marginal estimate of the ERF at $A_i^{(l)}$. Under the more typical
Bayesian framework, we could find \( \int_X \mu_x^{(k)}(a, x) \, dP(x) \) for each \( k = 1, 2, \ldots, K \). This would be equivalent to a Bayesian version of a G-computation estimator for the ERF (Keil et al. 2018). However, this process can be time-consuming when iterating for all \( K \) posterior samples instead of the \( L \) imputations, in addition to yielding a non-smooth ERF. The term on the left-hand side of the addition symbol in (4) is the residual error conditioned on \( A_i^{(l)} \) and \( X_i \). Since multiple imputation couples Bayesian methodology with regression techniques, it is necessary for each imputation that we approximate the variance for estimates of \( \theta(a) \) at each point \( a \in \mathcal{A} \), which is facilitated by the addition of this residual error term. Without the residual error component, the variance estimates within each imputation would be incorrect (Antonelli et al. 2020).

Local regression methods offer perhaps the most flexible means of projecting the pseudo-outcomes in (4) onto the support \( a \in \mathcal{A} \). Given a bandwidth \( h > 0 \), the ERF estimates we obtain for each imputation \( L = 1, 2, \ldots, L \) are denoted by \( \hat{\theta}_h^{(l)}(a) \). A pointwise estimate of the ERF is summarized by \( \bar{\theta}_h(a) = L^{-1} \sum_{l=1}^{L} \hat{\theta}_h^{(l)}(a) \). Because we regress each of the imputed pseudo-outcomes onto the support of the exposure, \( a \in \mathcal{A} \), the values \( \hat{\theta}_h^{(l)}(a) \) do not form a proper posterior of \( \theta(a) \). This result was noted also by Antonelli et al. (2020) who found that the posterior distribution of a regression-based estimator like \( \hat{\theta}_h^{(l)}(a) \) alone did not adequately characterize the uncertainty. To correct for this, our approach to estimation combines a kernel-weighted least-squares regression estimator similar to Kennedy et al. (2017) with the Bayesian approach of Antonelli et al. (2020). Instead of using a bootstrap estimator to find the empirical MCMC standard error used by Antonelli et al. (2020), we use an asymptotic standard error estimator derived by Kennedy et al. (2017). Details for estimating \( \hat{\theta}_h^{(l)}(a) \) and the accompanying standard errors for \( \bar{\theta}_h(a) \) using multiple imputation combining rules (Rubin 2004) are provided in Supplemental Section S2 along with the details on how to smooth the BART ERF estimates using kernel-weighted least-squares regression.

A simple regression calibration variation to the above approach would be to find a single imputation (i.e., \( L = 1 \)) of \( A_i \), let’s say \( \hat{A}_i \), then construct an estimator of the outcome mean \( \hat{\mu}_{\mathcal{Y}}(\hat{A}_i, X_i) \) and substitute that estimator into (4) replacing \( \mu_x^{(1)}(A_i^{(1)}, X_i) \). In this single imputation case, the ERF estimator would equal \( \hat{\theta}_h^{(1)}(a) \). We will see in the simulation that for a single imputation, the choice of \( \hat{A}_i \) is not so straightforward in the presence of measurement error, nor does such an approach adequately propagate the uncertainty created by measurement error without further correction. For more details on the local regression methods we apply, see Supplemental Section S2.

### 3.3. Issues Stemming from Congeniality

Our model contains two components, an imputation stage and an analysis stage. The imputation stage generates imputations of the latent exposures, \( A_i \) and \( S_{ij} \), while the analysis stage draws samples from the posterior distributions of the model parameters. Following the multiple imputation literature, the imputation component of our model must condition on the outcome to satisfy assumptions associated with congeniality, explained below. Congeniality states that the imputation and analysis stage models need to utilize the same data (Meng 1994), and violations of congeniality can bias parameter estimates. From a Bayesian
perspective, non-congeniality can arise due to cutting feedback or modularizing (Zigler et al. 2013) the components of a joint model. The limiting distribution from an MCMC of \( A_i^{(l)} \), and by extension the limiting distribution of the exposure effect, is ill-defined when cutting feedback between the outcome and exposure model in a Bayesian setting (Plummer 2015). Since we are necessarily using the imputations of \( A_i \) to fit the analysis model \( \mu_Y(A_i, X_i) \) and construct an estimate of the ERF, then to satisfy congeniality and avoid biasing the ERF estimate, the imputation models must be conditional on the outcome. This is accomplished in our method by using a fully Bayesian joint model-fitting scheme for the EPE, GPS, and outcome models.

However, sampling the latent exposures \( A_i \) conditioned on the outcome creates bi-directional feedback between the traditional “design stage” (GPS modeling) and analysis stage that are kept separate in most causal analyses. This seemingly defies research that emphasizes cutting feedback between the GPS and outcome model in a Bayesian causal analysis (Zigler et al. 2013). Notice that we include a model for the GPS in (2) that we fit in Sect. 3.1, yet we do not utilize the GPS in our outcome model nor our estimator of the ERF in Sect. 3.2. While the outcome data does not appear in the full conditionals for the GPS model parameters, it is used to generate new predictions of \( A_i \) which may indirectly create feedback affecting the GPS model estimates. To counteract the feedback problems created by congeniality, we removed the GPS adjustments from the doubly robust pseudo-outcome supported by Kennedy et al. (2017) (which appears in the Supplemental Section S4). These GPS adjustments appear to provide no utility in our context as demonstrated by a small simulation study contained within Supplemental Section S4. In this simulation, we draw particular attention to cases where the outcome model is misspecified but the GPS is correctly specified. In this scenario, there is almost no difference between using the GPS adjusted pseudo-outcome supported by Kennedy et al. (2017), and the pseudo-outcome we suppose in (4). We contend that this null result is attributable to the feedback created by the need to use congenial models for the imputation and analysis stages of the MCMC sampler.

4. NUMERICAL EXAMPLE

4.1. SIMULATION DESIGN

In this simulation study, we examine the performance of the method described in Sect. 3 in the presence and absence of exposure measurement error. In addition, we will evaluate the effects of model misspecification across the three components of the likelihood in (2), all while examining whether uncertainty is properly propagated into the final ERF estimate.

We test four different methods for measurement error correction. The first method naïvely ignores any measurement error and assumes \( \tilde{Z}_i \) is the true cluster-level exposure. The second method examines an extension to the regression calibration approach proposed by Wu et al. (2019), which we adapted to consider a continuous exposure, that corrects for prediction error but disregards the remaining classical error in (1) created by the remaining aggregation error. After finding predictions \( \hat{S}_{ij} \) for \( S_{ij} \) using least-squares regression conditioned on \( \tilde{S}_{ij} \) and \( W_{ij} \), a single imputation of \( A_i \) is produced by computing \( \hat{Z}_i = \hat{M}_i^{-1} \sum_{j=1}^{M_i} \hat{S}_{ij} \). We will refer to these two implementations that use either \( \tilde{Z}_i \) or \( \hat{Z}_i \) as the “single imputation”
approaches. The third and fourth methods that we examine are two “multiple imputation” approaches that follow the proposal in Sect. 3, with different outcome model specifications. In one of the multiple imputation variants, we let $\mu(l) \bar{Y}(A(l)i, X_i)$ be a BART model. In another, we let $\mu(l) \bar{Y}(A(l)i, X_i)$ be a correctly specified log-linear Poisson model, assuming unknown coefficient values (see Supplemental Sections S3 and S4 for details). To highlight the impacts of correcting measurement error and propagating uncertainty, we utilize estimation approaches that are as similar as possible across the four competing methods aside from how they correct for measurement error. For the single imputation approaches, we construct the pseudo-outcome in (4) in the exact same manner as in the multiple imputation cases, but for a single iteration, i.e., setting $L = 1$. For the single imputation methods, this means using the $\tilde{Z}_i$ or $\hat{Z}_i$ in place of $A(l)i$ and either $\hat{\mu}(\cdot)$ or $\tilde{\mu}(\cdot)$ in place of $\mu(l)(\cdot)$, respectively.

We also assume that $\phi_i = 0$ for all $i = 1, 2, \ldots, n$ (i.e., $A_i$ is independent), both in the data generation procedure and in the models we fit. In Supplemental Section S3, we describe the various data-generating mechanisms used to construct the simulation scenarios. For each scenario, we ran 500 iterations, in each one applying each of the four methods described above. In each iteration, we obtain pointwise estimates of $\theta(a)$ at 201 equally spaced exposure values across the range $a \in [6, 14]$. We report the relative bias, the residual mean square error, and coverage probabilities averaged across all the 500 iterations and all the 201 pointwise ERF estimates. We also report these same measurements for predictions at a single exposure level, $a = 11$, averaged across the 500 iterations. Briefly, we vary $n \in \{400, 800\}$, $m \in \{5n, 10n\}$, $\tau^2 \in \{0, 1, 2\}$, and $\omega^2 \in \{0, 1, 2\}$. Note that when $\tau^2 = 0$ the prediction error is absent, whereas when $\omega^2 = 0$ the remaining classical error is absent and $U_{ij}$ is limited to the prediction error. We also vary the degree of misspecification in the component models contained in (2) (Kang and Schafer 2007).

4.2. SIMULATION RESULTS

The main results of this simulation illustrate how adding measurement error can influence estimates of the ERF. Figure 3 shows the average ERF estimated with each of the four methods in Sect. 4.1 (relative to the true ERF) under simulation scenarios with no measurement error, prediction error only, classical error only, and both sources of error present. These results demonstrate that generating unbiased imputations of the exposure of interest is critical, whether using a single or multiple imputations. Note that from the description of the scenarios, we have $E[\tilde{S}_{ij}|A_i] = A_i$ when the EPE model is correctly specified. Therefore, the naïve approach and the regression calibration approach, which corrects for prediction error, produce similar results as shown by the root-mean-squared errors displayed in Fig. 3. This is because the prediction error is Berkson, so using regression calibration has a negligible effect. Also, note that the regression calibration and multiple imputation approaches perform similarly in scenarios without classical error (Fig. 3, top row), but the performance of regression calibration deteriorates when we simulate measurement error with aggregation error present. Meanwhile, the multiple imputation approaches retain accuracy (Fig. 3, bottom row). Some bias still lingers in each method, particularly at the extremes of the exposure
support (i.e., $a = 6$ and $a = 14$). This is typical of local regression methods where there is an implicit bias-variance tradeoff through the choice of $h$ (which is set to $h = 0.2$ when $n = 800$ and $h = 0.4$ when $n = 400$) (Hastie et al. 2001). Unsurprisingly, the accuracy of the different methods we test improves as both $n$ and $m$ increase (Table 1).

We can also see in Table 2 that the coverage probabilities from the 95% confidence interval estimates among the multiple imputation approaches offer a marked improvement over the alternatives. When using a correctly specified log-linear outcome model within a multiple imputation setting, we achieve coverage probabilities that match the nominal 95% confidence level. However, the coverage probabilities are imperfect when using a BART outcome model. BARTs’ tendency to under-report the dispersion of point estimates is well-documented (Wendling et al. 2018; Hahn et al. 2020; Nethery et al. 2021). The BART output also seems to produce biased estimates at several points along the curve, contributing to the low coverage probabilities. Despite this limitation, BART remains the preferred outcome
Table 1. Relative bias of the fitted ERFs averaged over the simulated ERF estimates when measurement error is present. The values in parentheses represent the statistics evaluated at \( a = 11 \). The check-marks indicates whether the corresponding model labeled in the column header is misspecified. The GLM approach refers to the multiple imputation implementation using a log-linear outcome model. The BART approach to multiple imputation uses a BART outcome model.

| \( n \) | \( m \) | \( \omega^2 \) | \( \tau^2 \) | GPS | Outcome | EPE | No Correction | Regression Calibration | BART Multiple Imputation | GLM Multiple Imputation |
|-------|-------|--------|----------|-----|---------|-----|----------------|------------------------|------------------------|------------------------|
| 400   | 2000  | 1      | 1        | 0.35 (−0.09) | 0.28 (−0.09) | 0.21 (−0.06) | −0.03 (−0.02) |
| 400   | 2000  | 1      | 2        | 0.42 (−0.11) | 0.30 (−0.09) | 0.28 (−0.06) | −0.05 (−0.02) |
| 400   | 2000  | 2      | 1        | 0.38 (−0.12) | 0.32 (−0.11) | 0.22 (−0.08) | −0.03 (−0.02) |
| 400   | 2000  | 2      | 2        | 0.42 (−0.14) | 0.29 (−0.13) | 0.21 (−0.08) | −0.05 (−0.02) |
| 400   | 4000  | 1      | 1        | 0.30 (−0.06) | 0.21 (−0.06) | 0.20 (−0.04) | −0.03 (−0.02) |
| 400   | 4000  | 1      | 2        | 0.31 (−0.08) | 0.18 (−0.07) | 0.21 (−0.05) | −0.06 (−0.02) |
| 400   | 4000  | 2      | 1        | 0.29 (−0.09) | 0.23 (−0.08) | 0.19 (−0.06) | −0.02 (−0.02) |
| 400   | 4000  | 2      | 2        | 0.30 (−0.11) | 0.19 (−0.10) | 0.15 (−0.07) | −0.04 (−0.02) |
| 800   | 4000  | 1      | 1        | 0.29 (−0.08) | 0.21 (−0.07) | 0.12 (−0.03) | −0.04 (0.00) |
| 800   | 4000  | 1      | 2        | 0.35 (−0.11) | 0.22 (−0.09) | 0.12 (−0.04) | −0.07 (0.00) |
| 800   | 4000  | 2      | 1        | 0.36 (−0.10) | 0.29 (−0.09) | 0.15 (−0.04) | −0.04 (0.00) |
| 800   | 4000  | 2      | 2        | 0.43 (−0.12) | 0.31 (−0.10) | 0.13 (−0.04) | −0.06 (0.00) |
| 800   | 8000  | 1      | 1        | 0.20 (−0.06) | 0.15 (−0.05) | 0.10 (−0.03) | −0.04 (−0.01) |
| 800   | 8000  | 1      | 2        | 0.26 (−0.06) | 0.16 (−0.05) | 0.10 (−0.03) | −0.07 (0.00) |
| 800   | 8000  | 2      | 1        | 0.22 (−0.08) | 0.17 (−0.07) | 0.10 (−0.03) | −0.04 (−0.01) |
| 800   | 8000  | 2      | 2        | 0.28 (−0.08) | 0.19 (−0.07) | 0.11 (−0.03) | −0.07 (−0.01) |
| 800   | 4000  | 2      | 1        | ✓ | 0.91 (−0.09) | 0.30 (−0.09) | 0.15 (−0.04) | −0.03 (0.00) |
| 800   | 4000  | 2      | 1        | ✓ ✓ | 0.18 (−0.20) | 0.11 (−0.20) | −0.01 (−0.14) | 0.04 (0.00) |
| 800   | 4000  | 2      | 1        | ✓ ✓ | 0.60 (−0.22) | 0.11 (−0.20) | −0.01 (−0.14) | 0.05 (0.00) |
| 800   | 4000  | 2      | 1        | ✓ ✓ | 0.52 (−0.06) | 0.46 (−0.05) | 0.29 (−0.01) | −0.03 (0.00) |
| 800   | 4000  | 2      | 1        | ✓ ✓ | 1.28 (0.02) | 0.59 (−0.02) | 0.40 (0.02) | −0.03 (0.00) |
| 800   | 4000  | 2      | 1        | ✓ ✓ | 0.15 (−0.21) | 0.08 (−0.20) | −0.05 (−0.15) | −0.09 (−0.05) |
| 800   | 4000  | 2      | 1        | ✓ ✓ | 0.52 (−0.25) | 0.02 (−0.23) | −0.09 (−0.16) | −0.09 (−0.05) |

model due to its strong predictive abilities even when the correct model specification is unknown (see the following paragraph). Since we average pointwise evaluations over the support \( A \), we can also see how the coverage probabilities are affected by using local regression techniques. This can be seen by examining the single pointwise evaluations at \( a = 11 \), which are closer to the nominal 95% confidence level than the average coverage probability over the 201 pointwise evaluations.

In the second part to our simulation study, where we induce model misspecification, we observe in Tables 1 and 2 that the ERF estimates are largely unaffected by outcome model misspecification alone with a BART model—remaining relatively unbiased as long as the GPS is still correctly specified. The same result is obviously not true for a parametric log-linear outcome model. Even though BART is supplied the untransformed covariates when model misspecification is present, it is quite adept at identifying complex nonlinear and interacting effects, making outcome model misspecification a moot consideration, so long as Assumptions 3 and 4 hold. That said, the coverage probability decreases considerably in
Table 2. Coverage probabilities averaged over the simulated ERF estimates when measurement error is present. The values in parentheses represent the statistics evaluated at $a = 11$. The check-marks indicates whether the corresponding model labeled in the column header is misspecified. Like in Table 1, the GLM approach refers to the multiple imputation implementation using a log-linear outcome model while the BART approach to multiple imputation implements a BART outcome model.

| $n$  | $m$  | $\omega^2$ | $e^2$ | GPS | Outcome | EPE | No Correction | Regression Calibration | BART Multiple Imputation | GLM Multiple Imputation |
|------|------|------------|------|-----|---------|------|---------------|-------------------------|-------------------------|-------------------------|
| 400  | 2000 | 1         | 1    | 0.44 (0.52) | 0.53 (0.54) | 0.75 (0.78) | 0.83 (0.90) |
| 400  | 2000 | 1         | 2    | 0.39 (0.40) | 0.52 (0.46) | 0.74 (0.73) | 0.70 (0.91) |
| 400  | 2000 | 2         | 1    | 0.39 (0.45) | 0.45 (0.48) | 0.71 (0.70) | 0.85 (0.90) |
| 400  | 2000 | 2         | 2    | 0.36 (0.35) | 0.45 (0.44) | 0.71 (0.62) | 0.75 (0.92) |
| 400  | 4000 | 1         | 1    | 0.53 (0.66) | 0.63 (0.69) | 0.78 (0.78) | 0.80 (0.91) |
| 400  | 4000 | 1         | 2    | 0.47 (0.58) | 0.63 (0.65) | 0.78 (0.76) | 0.64 (0.92) |
| 400  | 4000 | 2         | 1    | 0.47 (0.58) | 0.55 (0.64) | 0.74 (0.72) | 0.86 (0.92) |
| 400  | 4000 | 2         | 2    | 0.42 (0.48) | 0.58 (0.52) | 0.76 (0.71) | 0.69 (0.93) |
| 800  | 4000 | 1         | 1    | 0.42 (0.45) | 0.53 (0.52) | 0.76 (0.75) | 0.77 (0.96) |
| 800  | 4000 | 1         | 2    | 0.37 (0.38) | 0.50 (0.44) | 0.72 (0.74) | 0.62 (0.96) |
| 800  | 4000 | 2         | 1    | 0.37 (0.38) | 0.41 (0.39) | 0.73 (0.71) | 0.82 (0.96) |
| 800  | 4000 | 2         | 2    | 0.35 (0.30) | 0.43 (0.38) | 0.72 (0.72) | 0.65 (0.94) |
| 800  | 8000 | 1         | 1    | 0.50 (0.67) | 0.65 (0.69) | 0.78 (0.83) | 0.74 (0.94) |
| 800  | 8000 | 1         | 2    | 0.44 (0.56) | 0.59 (0.60) | 0.72 (0.76) | 0.59 (0.94) |
| 800  | 8000 | 2         | 1    | 0.45 (0.52) | 0.55 (0.58) | 0.78 (0.78) | 0.78 (0.94) |
| 800  | 8000 | 2         | 2    | 0.42 (0.45) | 0.56 (0.53) | 0.73 (0.74) | 0.62 (0.94) |
| 800  | 4000 | 2         | 1    | ✓    | 0.30 (0.47) | 0.42 (0.40) | 0.71 (0.73) | 0.76 (0.93) |
| 800  | 4000 | 2         | 2    | ✓    | 0.25 (0.02) | 0.32 (0.03) | 0.55 (0.14) | 0.32 (0.96) |
| 800  | 4000 | 2         | 2    | ✓    | ✓    | 0.19 (0.02) | 0.30 (0.02) | 0.55 (0.15) | 0.32 (0.97) |
| 800  | 4000 | 2         | 2    | ✓    | ✓    | 0.45 (0.54) | 0.49 (0.58) | 0.75 (0.82) | 0.75 (0.93) |
| 800  | 4000 | 2         | 2    | ✓    | ✓    | 0.34 (0.66) | 0.49 (0.61) | 0.74 (0.82) | 0.74 (0.93) |
| 800  | 4000 | 2         | 2    | ✓    | ✓    | ✓    | 0.18 (0.01) | 0.28 (0.01) | 0.44 (0.08) | 0.37 (0.77) |
| 800  | 4000 | 2         | 2    | ✓    | ✓    | ✓    | 0.15 (0.00) | 0.25 (0.00) | 0.41 (0.04) | 0.38 (0.72) |

scenarios with a misspecified outcome model relative to the results of the scenario where all models are correctly specified.

When the GPS model is misspecified, the level of bias increases substantially in each of the ERF estimates, both from the single imputation and multiple imputation models. A perplexing exception to this result is when both the GPS and the outcome model are misspecified, in which case the bias returns to the nominal levels observed in cases when the GPS is correctly specified and a BART outcome model is used. Given these results, and the results of the simulation in Supplemental Section S4, it is evident that finding an approximately correct GPS model may be even more important than finding a correct outcome model as we originally suggested. Doing so might decrease the precision of the imputations for $A_i$. Finally, when the EPE model is “misspecified”, meaning $\mathbb{E}(\tilde{S}_{ij}|A_i) \neq A_i$ implying $\mathbb{E}(\tilde{Z}_i|A_i) \neq A_i$, then we see the most severe levels of bias using the naïve (no correction) approach. Since the multiple imputation and regression calibration approaches correctly model $S^{(l)}_{ij}$ and $\hat{S}_{ij}$, respectively, such that $\mathbb{E}(S^{(l)}_{ij}|A_i) = A_i$ and $\mathbb{E}(\hat{S}_{ij}|A_i) = A_i$, this prediction bias does not affect the estimates of the exposure response curves using these two approaches.
5. APPLIED EXAMPLE

Recall the example described in Sect. 1.1. Wu et al. (2019) used regression calibration methods on the same data to correct for measurement error for the grid-year PM$_{2.5}$ predictions. Subsequently, they aggregated the corrected grid-year predictions to the ZIP-code-year level, categorized the imputed exposures based on standard EPA cutoffs, and examined a variety of GPS-based implementations including matching, sub-classification, and weighting to estimate the effect of PM$_{2.5}$ on mortality. Their results suggest that increasing PM$_{2.5}$ exposure led to excess mortality events.

In this analysis, we re-examine the same data as Wu et al. (2019), applying the proposed multiple imputation approach to estimate the ERF relating PM$_{2.5}$ with all-cause mortality. For ZIP-code-year $i$, the exposure $A_i$ is the annual average PM$_{2.5}$ concentration (in $\mu g/m^3$), the outcome $Y_i$ is the number of deaths observed, and $N_i$ is the person-years at risk. We have exposure data on $n = 28,626$ ZIP-code-years in New England covered by $m = 2,395,588$ 1km $\times$ 1km grid-years, each with error-prone PM$_{2.5}$ exposure predictions $\tilde{S}_{ij}$. A subset of 83 grids within 75 ZIP-codes have error-free PM$_{2.5}$ measurements from monitoring stations in years 2000–2012 (Fig. 1). While it is plausible that the $\tilde{S}_{ij}$ and $S_{ij}$ within ZIP-code-years are correlated, the majority of this correlation can be accounted for simply by conditioning on $A_i$, thus enabling Assumptions 1 and 2.

For the multiple imputation approach, we assume the outcome model in (3). As in the simulation study, we implement two different single imputation approaches. In the first approach, we ignore prediction error altogether and suppose $\tilde{Z}_i$ is the true exposure (we call this the “no correction” approach). In the second approach, we use least squares regression on the validation data to get the regression calibrated $\hat{Z}_i$. We then use these single imputations of the exposure in a BART model to construct an estimate of the mean function, either $\tilde{\mu} \tilde{Y}(\tilde{Z}_i, X_i)$ or $\hat{\mu} \tilde{Y}(\hat{Z}_i, X_i)$. For both the single ($L = 1$) and multiple imputation approaches ($L = 100$), local regression was applied according to Sect. 3.2. We collected 5,000 MCMC samples, thinned by every 50 iterations, after a burn-in of another 5,000 iterations. Diffuse priors specified in Sect. 3.1 were used for the GPS model parameters, the EPE model parameters, as well as for the BART model. As opposed to the simulation study, we account for spatial correlation between the ZIP-code exposures by sampling $\phi_i$ in (2). In this example, the matrix $V$ is a binary adjacency matrix with each row and column representing a different ZIP-code-year.

Figure 1 provides a heat map displaying the difference between the naïve exposure imputations of $\tilde{Z}_i$ and the posterior means $L^{-1} \sum_{l=1}^{L} A_i^{(l)}$ for each ZIP code in 2010. Here we can see substantial differences in the imputed exposure values in some areas of New England. Figure 4 shows the ERF estimates from each of the three methods we examined. In the curve estimated with multiple imputation, we can see a modest increase in the rate of all-cause mortality from 4.7% when annual average PM$_{2.5}$ is at 5 $\mu g/m^3$ to 5.0% when annual average PM$_{2.5}$ is at 15 $\mu g/m^3$. Relative to the uncorrected and regression calibration curves, the curve estimated with multiple imputation has a steeper gradient at lower levels of PM$_{2.5}$. The steepest increase in mortality occurs when PM$_{2.5}$ changes from about 5
Estimating a Causal Exposure Response

Figure 4. ERF estimate of PM$_{2.5}$ on all-cause mortality in New England between 2000 and 2012 amongst Medicare recipients under three different approaches to measurement error correction. The grey ribbon represents the 95% confidence interval computed from our multiple imputation approach. The histogram underlying the curves corresponds with the empirical distribution of the EPEs (Color figure online).

μg/m$^3$ to 10 μg/m$^3$, which is policy-relevant given the World Health Organization’s recent decision to lower the recommended limit of annual average PM$_{2.5}$ to 5 μg/m$^3$ (World Health Organization 2021).

We suspect that the minor differences between the three curves are primarily attributable to the small variation of the classical measurement error, ω$^2$, which we found to be equal to about 0.0187 (95% CI 0.0183–0.0189). The posterior mean for the conditional prediction error variance, τ$^2$, on the other hand was 1.340 (95% CI 1.212, 1.446). However, as we demonstrated in the simulation study, correcting for the prediction error is secondary to correcting for the classical measurement including the aggregation error, except when E($\tilde{S}_{ij}|A_i)$ ≠ $A_i$. This reasoning is further cemented in this illustrative example—we can see that E($\tilde{S}_{ij}|A_i)$ ≈ $A_i$, and so the difference between the curves with no measurement error correction and regression calibration is relatively small.

More generally, an interesting feature of the estimated curves is the slight dip in all-cause mortality when PM$_{2.5}$ is around 11 μg/m$^3$, which seems to suggest that increasing PM$_{2.5}$ from 11 μg/m$^3$ to 12.5 μg/m$^3$ is protective. In reality, this phenomenon is probably a result of slight over-fitting (more specifically under-smoothing of the curves). This is evidenced by the fact that there are fewer ZIP codes experiencing higher PM$_{2.5}$ levels, resulting in wider confidence bands at this end of the exposure response curve and suggesting that there
is substantial uncertainty about the shape of the curve in this region. In general, when there is a higher density of PM$_{2.5}$ measurements (see the histogram in Fig. 4), the more precise the exposure response curve estimates become. When the concentration of PM$_{2.5}$ measurements is lower, then any highly leveraged outliers might exert strong influence on the shape of the curve.

6. DISCUSSION

We developed a multiple imputation framework that addresses exposure measurement error when estimating a causal ERF. We adapted a Bayesian approach for correcting measurement error to generate imputations of the true exposures assuming non-differential classical measurement error. We then constructed a flexible representation of the ERF using kernel smoothed output from a Bayesian additive regression tree. These two steps interface with one another over several iterations. Since the final estimator isn’t formally Bayesian due to post-processing the BART-adjusted pseudo-outcome with kernel-weighted least squares regression, we regard it as a multiple imputation-based estimator. We then described some issues that we encountered with this approach regarding the seeming conflict that exists between cutting feedback and appeasing rules about congeniality. We provided several simulated numerical examples that showcase why correcting for exposure measurement error within a causal-analysis is important to obtain unbiased estimates and valid inferences.

In the real data example, we estimated the exposure response function associating all-cause mortality with annual average PM$_{2.5}$ exposures. At first glance, the three curves we estimate with varying degrees of measurement error correction all seem fairly similar for this particular analysis. These results seem to indicate that there is little benefit to correcting for measurement error while estimating ERFs, though the attenuation at lower PM$_{2.5}$ levels is meaningful. However, the reason for the similarities is mainly due to the low levels of measurement error. As measurement error increases, so does the attenuation bias. Indeed, our attempts to correct for measurement error in our applied example do not seem to have a large impact on the results associating PM$_{2.5}$ with mortality. Therefore, this analysis helps validate other analyses that used similar data but did not correct for measurement error.

While Medicare data limitations necessitate use of ZIP-code aggregate exposures in our analysis, there are various simplifications to Sects. 2 and 3 that can be made to accommodate scenarios where certain measurement error sources are absent. If we were provided individual-level outcomes and error-prone exposure data, then the measurement error would be closer to a Berkson-type rather than classical (Bateson and Wright 2010). Accounting for measurement error in this problem would require a different EPE model than the one proposed; however, adapting our framework should be straightforward given the adaptability of the Bayesian framework that we outlined (Carroll et al. 2006). Lastly, our approach assumes non-differential measurement error, which in our context means that $Y_i$ is unaffected by $\tilde{S}_{ij}$ given $A_i$. More work is required to generalize this methodology to be useful in cases where the measurement error is differential.

Another line of future work involves relaxing the independence and homoscedastic assumptions surrounding the measurement error (Assumptions 1 and 2). Since these mod-
els are deployed within an MCMC framework, more sophisticated spatiotemporal modeling techniques might be employed to better understand the measurement error residing in the data. Second, alternatives to the parametric GPS and EPE models were not implemented in this work. However, more flexible nonparametric methods may help reduce model-dependency of the ERF. We relaxed some of this model-dependency by using a BART outcome model. However, the BART implementation we employed assumes the outcome is approximately Gaussian. There is an option for fitting log-linear BART models (Murray 2021), which might ameliorate some of the problems we encountered with BART, like the poor coverage probabilities observed in the simulation study (Hahn et al. 2020). However, at the time of writing, code to fit this alternative BART model could not be obtained. This problem may also be addressed by choosing a different outcome model as demonstrated in the simulations.

REPRODUCIBLE R CODE

The R functions used to fit the model in Sect. 3 along with the code used to run the simulation studies in Sects. 4 and S4 are available at https://github.com/kevjosey/causal-me/.

ACKNOWLEDGEMENTS

Funding was provided by the National Institute of Health (NIH) grants (T32 ES007142, R01 ES030616, R01 ES028033, R01 ES026217, K01 ES032458-01) and HEI grant 4953-RFA14-3/16-4. The contents are solely the responsibility of the grantee and do not necessarily represent the official views of the funding agencies. Further, the funding agencies do not endorse the purchase of any commercial products or services related to this publication.

REFERENCES

Anderson HR (2009) Air pollution and mortality: a history. Atmos Environ 43(1):142–152
Antonelli J, Papadogeorgou G, Dominici F (2020) Causal inference in high dimensions: a marriage between bayesian modeling and good frequentist properties Biometrics (In Press)
Bateson TF, Wright JM (2010) Regression calibration for classical exposure measurement error in environmental epidemiology studies using multiple local surrogate exposures. Am J Epidemiol 172(3):344–352
Braun D, Gorfine M, Parmigiani G, Arvold ND, Dominici F, Zigler C (2017) Propensity scores with misclassified treatment assignment: a likelihood-based adjustment. Biostatistics 18(4):695–710
Brook RD, Rajagopalan S, Pope CA III, Brook JR, Bhatnagar A, Diez-Roux AV, Holguin F, Hong Y, Luepker RV, Mittleman MA et al (2010) Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American heart association. Circulation 121(21):2331–2378
Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC, Boca Raton, FL, USA
Chipman HA, George EI, McCulloch RE (2010) Bart: Bayesian additive regression trees. Ann Appl Statistics 4(1):266–298

[Received September 2021. Revised July 2022. Accepted July 2022. Published Online September 2022.]
Cole SR, Chu H, Greenland S (2006) Multiple-imputation for measurement-error correction. Int J Epidemiol 35(4):1074–1081

Danesh Yazdi M, Wang Y, Di Q, Wei Y, Requia WJ, Shi L, Sabath MB, Dominici F, Coull BA, Evans JS et al (2021) Long-term association of air pollution and hospital admissions among medicare participants using a doubly robust additive model. Circulation 143(16):1584–1596

Di Q, Koutrakis P, Schwartz J (2016) A hybrid prediction model for pm2.5 mass and components using a chemical transport model and land use regression. Atmos Environ 131:390–399

Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Choirat C, Dominici F, Schwartz JD (2017) Air pollution and mortality in the medicare population. N Engl J Med 376(26):2513–2522

Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, Samet JM (2006) Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA 295(10):1127–1134

Gill RD, Robins JM (2001) Causal inference for complex longitudinal data: the continuous case. Ann Statist 29(6):1785–1811

Haber G, Sampson J, Graubard B (2020) Bias due to Berkson error: issues when using predicted values in place of observed covariates. Biostatistics 22(4):858–872

Hahn PR, Murray JS, Carvalho CM (2020) Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). Bayesian Anal 15(3):965–1056

Hajat S, Anderson H, Atkinson R, Haines A (2002) Effects of air pollution on general practitioner consultations for upper respiratory diseases in London. Occup Environ Med 59(5):294–299

Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer Series in Statistics. Springer New York, NY, USA

Kang JDY, Schafer JL (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci 22(4):523–539

Keil AP, Daza EJ, Engel SM, Buckley JP, Edwards JK (2018) A Bayesian approach to the g-formula. Stat Methods Med Res 27(10):3183–3204

Kennedy EH, Ma Z, McHugh MD, Small DS (2017) Nonparametric methods for doubly robust estimation of continuous treatment effects. J R Stat Soc Ser B Stat Methodol 79(4):1229

Kim W, Kwon K, Kwon S, Lee S (2018) The identification power of smoothness assumptions in models with counterfactual outcomes. Quant Econ 9(2):617–642

Kioumourtzoglou MA, Spiegelman D, Szpiro AA, Sheppard L, Kaufman JD, Yanosky JD, Williams R, Laden F, Hong B, Suh H (2014) Exposure measurement error in 2.5 health effects studies: a pooled analysis of eight personal exposure validation studies. Environ Health 13(1):1–11

Lee D (2013) Carbayes: an R package for bayesian spatial modeling with conditional autoregressive priors. J Stat Softw 55(13):1–24

Lenis D, Ebnesajjad CF, Stuart EA (2017) A doubly robust estimator for the average treatment effect in the context of a mean-reverting measurement error. Biostatistics 18(2):325–337

Leroux BG, Lei X, Breslow N (2000) Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Halloran ME, Berry D (eds) Statistical models in epidemiology, the environment, and clinical trials. The IMA volumes in mathematics and its applications, vol. 116, pp 179–191. Springer, New York

Lewbel A (2007) Estimation of average treatment effects with misspecification. Econometrica 75(2):537–551

Meng X-L (1994) Multiple-imputation inferences with uncongenial sources of input. Stat Sci 9(4):538–558

Murray JS (2021) Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. J Am Stat Assoc 116(534):756–769

Nethery RC, Mealli F, Sacks JD, Dominici F (2021) Evaluation of the health impacts of the 1990 clean air act amendments using causal inference and machine learning. J Am Stat Assoc 116(535):1128–1139

Papadogeorgou G, Mealli F, Zigler CM (2019) Causal inference with interfering units for cluster and population level treatment allocation programs. Biometrics 75(3):778–787

Plummer M (2015) Cuts in Bayesian graphical models. Stat Comput 25(1):37–43
Pope CA III, Turner MC, Burnett RT, Jerrett M, Gapstur SM, Diver WR, Krewski D, Brook RD (2015) Relationships between fine particulate air pollution, cardiometabolic disorders, and cardiovascular mortality. Circ Res 116(1):108–115

Ren B, Wu X, Braun D, Pillai N, Dominici F (2021) Bayesian modeling for exposure response curve via gaussian processes: causal effects of exposure to air pollution on health outcomes arXiv preprint arXiv:2105.03454

Rhee J, Dominici F, Zanobetti A, Schwartz J, Wang Y, Di Q, Balmes J, Christiani DC (2019) Impact of long-term exposures to ambient pm2.5 and ozone on ards risk for older adults in the united states. Chest 156(1):71–79

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688

Rubin DB (2004) Multiple imputation for nonresponse in surveys. Wiley, New York, NY, USA

Webb-Vargas Y, Rudolph KE, Lenis D, Murakami P, Stuart EA (2017) An imputation-based solution to using mismeasured covariates in propensity score analysis. Stat Methods Med Res 26(4):1824–1837

Wendling T, Jung K, Callahan A, Schuler A, Shah N, Gallego B (2018) Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. Stat Med 37(23):3309–3324

World Health Organization (2021) New who global air quality guidelines aim to save millions of lives from air pollution https://www.who.int/

Wu X, Braun D, Kioumourtzoglou M-A, Choirat C, Di Q, Dominici F (2019) Causal inference in the context of an error prone exposure: air pollution and mortality. Ann Appl Statistics 13(1):520

Wu X, Braun D, Schwartz J, Kioumourtzoglou M, Dominici F (2020) Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly. Sci Adv 6(29):eaba5692

Wu X, Mealli F, Kioumourtzoglou MA, Dominici F, Braun D (2018) Matching on generalized propensity scores with continuous exposures arXiv:1812.06575

Yazdi MD, Wang Y, Di Q, Zanobetti A, Schwartz J (2019) Long-term exposure to pm2.5 and ozone and hospital admissions of medicare participants in the southeast usa. Environ Int 130:104879

Zanobetti A, Franklin M, Koutrakis P, Schwartz J (2009) Fine particulate air pollution and its components in association with cause-specific emergency admissions. Environ Health 8(1):1–12

Zhu L, Ge X, Chen Y, Zeng X, Pan W, Zhang X, Ben S, Yuan Q, Xin J, Shao W et al (2017) Short-term effects of ambient air pollution and childhood lower respiratory diseases. Sci Rep 7(1):1–7

Zigler CM, Watts K, Yeh RW, Wang Y, Coull BA, Dominici F (2013) Model feedback in Bayesian propensity score estimation. Biometrics 69(1):263–273

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.