The Annals of Applied Statistics
2019, Vol. 13, No. 2, 1242–1267
https://doi.org/10.1214/18-AOAS1231

ESTIMATING POPULATION AVERAGE CAUSAL EFFECTS IN THE PRESENCE OF NON-OVERLAP: THE EFFECT OF NATURAL GAS COMPRESSOR STATION EXPOSURE ON CANCER MORTALITY[1]

BY RACHEL C. NETHERY*, FABRIZIA MEALLI† AND FRANCESCA DOMINICI*

*Harvard T.H. Chan School of Public Health** *and University of Florence*†

Most causal inference studies rely on the assumption of overlap to estimate population or sample average causal effects. When data suffer from non-overlap, estimation of these estimands requires reliance on model specifications due to poor data support. All existing methods to address non-overlap, such as trimming or down-weighting data in regions of poor data support, change the estimand so that inference cannot be made on the sample or the underlying population. In environmental health research settings where study results are often intended to influence policy, population-level inference may be critical and changes in the estimand can diminish the impact of the study results, because estimates may not be representative of effects in the population of interest to policymakers. Researchers may be willing to make additional, minimal modeling assumptions in order to preserve the ability to estimate population average causal effects. We seek to make two contributions on this topic. First, we propose a flexible, data-driven definition of propensity score overlap and non-overlap regions. Second, we develop a novel Bayesian framework to estimate population average causal effects with minor model dependence and appropriately large uncertainties in the presence of non-overlap and causal effect heterogeneity. In this approach the tasks of estimating causal effects in the overlap and non-overlap regions are delegated to two distinct models suited to the degree of data support in each region. Tree ensembles are used to nonparametrically estimate individual causal effects in the overlap region, where the data can speak for themselves. In the non-overlap region where insufficient data support means reliance on model specification is necessary, individual causal effects are estimated by extrapolating trends from the overlap region via a spline model. The promising performance of our method is demonstrated in simulations. Finally, we utilize our method to perform a novel investigation of the causal effect of natural gas compressor station exposure on cancer outcomes. Code and data to implement the method and reproduce all simulations and analyses is available on Github (https://github.com/rachelnethery/overlap).

Received May 2018; revised November 2018.

[1]Support for this work was provided by NIH Grants 5T32ES007142-35, R01ES028033, P01CA134294, R01GM111339, R35CA197449, R01ES026217, P50MD010428, and R01MD012769. The authors also received support from EPA Grants 83615601 and 83587201-0, Health Effects Institute Grant 4953-RFA14-3/16-4, and the Dipartimenti Eccellenti 2018-2022 Italian ministerial funds.

Key words and phrases. Overlap, propensity score, Bayesian additive regression trees, splines, natural gas, cancer mortality.

1. Introduction.

1.1. *Natural gas compressor stations and cancer mortality.* During the last several decades, the United States (US) has witnessed a sharp increase in the incidence of thyroid cancer which now accounts for 1–1.5% of all newly diagnosed cancer cases (Pellegriti et al. (2013)). Increased exposure of the population to radiation and carcinogenic environmental pollutants is blamed, in part, for this increase.

During the last several decades, US natural gas (NG) production has also increased rapidly. NG production systems have recently received attention as a potential source of human exposure to carcinogens and endocrine-disrupting chemicals (Kassotis et al. (2016)). Recent epidemiological studies have found links between NG production and leukemia and between NG production and thyroid cancer (Finkel (2016), McKenzie et al. (2017)). The relationship between NG systems and thyroid cancer could be of particular interest due to their coincident rise.

While most previous studies have focused on the health effects of exposure to NG production sites (e.g., drilling wells) (Finkel (2016), McKenzie et al. (2017)), we turn our attention instead to the potential health effects of NG distribution systems. Specifically, we aim to provide the first data-driven epidemiological study of the causal effects of NG compressor station exposure on thyroid cancer and leukemia mortality rates.

NG compressor stations are pumping stations located at 40–70 mile intervals along NG pipelines. They keep pressure in the pipelines so that NG flows in the desired direction (Messersmith, Brockett and Loveland (2015)). The operations at compressor stations have raised health concerns for residents of nearby communities (Southwest Pennsylvania Environmental and Project (2015)). In this paper we exclude from consideration the health impacts of accidents at compressor stations and focus on the potentially harmful exposures to nearby communities resulting from the normal operations of compressor stations. Fugitive emissions, or unintended leaking of chemicals from the compressor station equipment, are known to occur but are not well characterized. NG compressor stations also routinely conduct "blowdowns," in which pipelines and equipment are vented to reduce pressure (Kloczko (2015)) and any chemicals present in the pipeline are reportedly released into the air in a 30–60 meter plume of gas (Southwest Pennsylvania Environmental and Project (2015)). Little is known about the specific types of chemicals emitted.

While airborne emissions from compressor stations are regulated by the EPA under the Clean Air Act (Messersmith, Brockett and Loveland (2015)), air quality studies in Pennsylvania and Texas have discovered harmful chemicals in excess of standards near NG compressor stations (Pennsylvania Department of Environmental Protection (2010), Wolf Eagle Environmental (2009)). These chemicals include methane, ethane, propane and numerous benzene compounds. Benzene is a known

carcinogen (Golding and Watson (1999), Maltoni et al. (1989)), and many of these compounds are known or suspected endocrine disruptors (US EPA (2018)).

Motivated by these findings we present an investigation of the causal effects of compressor stations on county-level thyroid cancer and leukemia mortality rates. From a sample of 978 counties from the mid-western region of the US, we obtained their NG compressor stations exposure status, their thyroid cancer and leukemia mortality rates and many suspected confounders of this relationship. While we would like to apply a classic nonparametric causal inference analysis rooted in the potential outcomes approach (Rubin (1974)), the data exhibit non-overlap, that is, in some areas of the confounder space there is little or no variability in the exposure status of the units. Due to this non-overlap, any attempt to adjust for confounding when estimating the population average causal effect must rely upon model-based extrapolation, because we have insufficient data to infer about missing potential outcomes in those regions of the confounder space. Thus, nonparametric causal inference methods may yield unreliable results.

In this paper we seek to make two methodological contributions to the causal inference literature. First, we introduce a flexible, data-driven definition of sample propensity score overlap and non-overlap regions. Second, we propose a novel approach to estimating population average causal effects in the presence of non-overlap. Using this approach, the sample is split into a region of overlap (RO) and a region of non-overlap (RN) and distinct models, appropriate for the amount of data support in each region, are developed and applied to estimate the causal effects in the two regions separately. We have found that the proposed approach leads to improved estimation of the population average causal effects compared to existing methods. Moreover, we apply this method to estimate the population average causal effect of compressor station exposure on thyroid cancer and leukemia mortality.

1.2. *Causal inference notation and assumptions.* We first introduce notation that will be used throughout this article. For subject i $(i = 1, \ldots N)$, Y_i^{obs} will denote the observed outcome (here, it will be assumed to be a continuous random variable; in Section 2.4 we introduce analogous notation for the binary outcomes setting), E_i will denote a binary treatment or exposure and $\boldsymbol{X}_i$ will denote a vector of observed confounders. Under the stable unit treatment value assumption (Rubin (1980)) potential outcomes $Y_i(1)$ and $Y_i(0)$, corresponding to the outcome that would be observed under scenarios $E_i = 1$ and $E_i = 0$ respectively, exist for each unit. Only one of these potential outcomes can be observed such that $Y_i^{\text{obs}} = E_i Y_i(1) + (1 - E_i) Y_i(0)$. We denote each unit's missing potential outcome as Y_i^{mis}, that is, $Y_i^{\text{mis}} = (1 - E_i) Y_i(1) + E_i Y_i(0)$. An individual causal effect refers to the difference in potential outcomes for a unit, that is, $\Delta_i = Y_i(1) - Y_i(0)$. The sample average causal effect is $\Delta_S = \frac{1}{N} \sum_{i=1}^{N} \Delta_i$, the population conditional average causal effect is $\Delta_{P|\boldsymbol{x}} = E[Y(1) - Y(0)|\boldsymbol{x}]$ and the population average causal effect is $\Delta_P = E_{\boldsymbol{X}}[E[Y(1) - Y(0)|\boldsymbol{X}]]$.

The identifiability of population level causal effects in observational studies relies upon the assumptions of (1) unconfoundedness and (2) positivity. Unconfoundedness implies that all confounders of the relationship between exposure and outcome are observed, that is, $E_i \perp\!\!\!\perp (Y_i(1), Y_i(0))|\boldsymbol{X}_i$. We assume throughout that unconfoundedness holds. We also assume positivity, stated mathematically as $0 < P(E_i = 1|\boldsymbol{X}_i) < 1$. This means that each unit in the population has positive probability of obtaining either exposure status.

We relax the assumption of overlap, closely related to positivity, which is required to nonparametrically estimate sample and population average causal effects. The term overlap refers to the overlap of the confounder distributions across the exposure groups. Non-overlap occurs when every unit in the population is eligible to receive either exposure, but, by chance, few or no units from one exposure group are observed in some confounder strata (Westreich and Cole (2010)). Non-overlap can be a population level feature or a finite sample issue only. Here, we address problems arising from finite sample non-overlap, that is, scenarios in which the population exhibits complete overlap but, in some areas of the confounder space, data are sparse for one or both exposure groups leading to representative samples with non-overlap.

In the presence of non-overlap, sample and population average causal effect estimates generally suffer from bias and increased variance unless they are able to rely on the additional assumption of correct model specification (King and Zeng (2005), Petersen et al. (2012)). The overlap assumption can be evaluated by comparing the empirical distribution of the estimated propensity score, $\hat{\xi}_i = \hat{P}(E_i = 1 \mid \boldsymbol{X})$, between the exposure groups (Austin (2011), Rosenbaum and Rubin (1983)), assuming the propensity score is well estimated. The further assumption that non-overlap is a finite sample feature only is generally untestable but can sometimes be evaluated using subject-matter expertise. While the assumption of unconfoundedness becomes more plausible as the number of covariates grows, the likelihood of non-overlap increases (Cole and Hernán (2008), D'Amour et al. (2017)); thus, non-overlap is an increasing problem in our era of high dimensional data.

1.3. *Existing methods for estimating causal effects in the presence of non-overlap*. Methods for reducing the bias and variance of causal effect estimates in the presence of propensity score non-overlap are abundant in the causal inference literature (Cole and Hernán (2008), Crump et al. (2009), Li, Morgan and Zaslavsky (2018), Petersen et al. (2012)); however, to our knowledge, all of the existing methods modify the estimand and its interpretation so that neither sample nor population average causal effect estimates can be obtained. In many medical and health research settings, such as evaluation of treatments, the aim of the research is to help clinicians choose between various forms of treatment for patients who are likely to adhere to any of the available treatments. In these contexts convenience samples are common and modified estimands may be equally informative

as or more informative than sample or population level estimands. However, in environmental health applications like the one considered here, study samples are often carefully selected to reflect a population of interest to policymakers. Then, the primary aim is to estimate the burden of disease attributable to certain contaminants for the whole sample or underlying population in order to ultimately inform regulatory policies. Another example is recent studies of the health effects of air pollution exposure in the Medicare population (Di et al. (2017a), Di et al. (2017b)).

The most commonly recommended approach for handling propensity score non-overlap is "trimming" or discarding observations in regions of poor data support (Crump et al. (2009), Ho et al. (2007), Petersen et al. (2012)). However, trimming allows only for the estimation of the average causal effect *in the trimmed sample*, and it changes the asymptotic properties of estimators in ways that are often overlooked (Yang and Ding (2018)). Recent developments in the context of weighting approaches to causal inference, such as the overlap weights of Li, Morgan and Zaslavsky (2018), Li and Thomas (2018), may provide more interpretable estimands than trimming.

In contrast to the existing literature, which emphasizes the removal or down-weighting of data in regions of poor support, we propose a method that (1) minimizes model dependence where possible and (2) performs model-based extrapolation in a principled manner where necessary yielding estimates of the population level estimand with small bias and appropriately large uncertainty. This method will be most valuable in environmental health and other applications where preserving population-level inference in spite of non-overlap is critical.

Our Bayesian modeling approach estimates individual causal effects (Δ_i) in the RO and the RN separately. In the first stage of this procedure, a nonparametric Bayesian Additive Regression Tree (BART) (Chipman, George and McCulloch (2010)) is fit to the data in the RO to estimate causal effects Δ_i for each observation in the RO where data support is abundant. In the second stage a spline (SPL) is fitted to the estimated Δ_i in the RO to capture trends in the causal effect surface. The SPL is used to extrapolate those trends to estimate Δ_i for observations in the RN where insufficient data support requires reliance on model specifications/extrapolation. The data in the RN are excluded from all model fitting, so that the models are not influenced by data-sparse regions; however, after model fitting to data in the RO, the observed potential outcomes in the RN are employed as covariate values to aid in prediction of causal effects in the RN, so that we maximally leverage the information in the RN. Because of the flexibility of both BART and SPL, our method captures nonlinearities and causal effect heterogeneity.

In Section 2 we provide a data-driven definition of the RO and RN, and we introduce our method, called BART + SPL. Simulations in Section 3 demonstrate how BART + SPL can yield improved population average causal effect estimates relative to existing methods and provide guidance in specifying tuning parameters. In Section 4 BART + SPL is applied to estimate the effect of NG compressor station exposure on thyroid cancer and leukemia mortality rates. We conclude with a summary of our findings in Section 5.

2. Methods.

2.1. *Definition of overlap and non-overlap regions.* King and Zeng (2005) proposed the use of the convex hull of the data to define RO and RN; however, this criteria is very conservative. Crump et al. (2009) offer a definition of the RO with the goal of identifying a region of the data that will produce a minimum variance average causal effect estimate. Their definition may be optimal in the context of trimming in conjunction with nonparametric causal estimators, but it is not a general definition of the region where strong data support is observed. BART itself has also been proposed as a method for identifying the RN (Hill and Su (2013)).

We intend to provide a more general characterization of the RO and RN here. We use the estimated propensity scores, $\hat{\xi}_i$, to define the RO, O, and the RN, $O^{\perp}$, in the sample. Throughout this section we assume that the propensity score model has been correctly specified (or that the true propensity score is known, in which case ξ_i can be substituted for $\hat{\xi}_i$ in the following); however, we demonstrate and discuss the performance of our method under propensity score model misspecification using simulations in Section 3. Let $\hat{\xi}_{(j)}$ denote the jth order statistic of the $\hat{\xi}_i$ and $P = [\hat{\xi}_{(1)}, \hat{\xi}_{(N)}]$ be the subspace of $(0, 1)$ over which the $\hat{\xi}_i$ are observed.

Our definition allows every point in P to be assigned to either O or $O^{\perp}$. The user must prespecify two parameters, denoted a and b, which are used to identify O. $O^{\perp}$ is then defined as the complement of O relative to P. Consider any point $o \in P$. The idea behind our definition of overlap is that, if more than b units from each exposure group have estimated propensity scores lying within some open interval of size a covering o, then o is included in the region of overlap. Thus, a is an interval length, that is, a portion of the range of the estimated propensity score, and b is a portion of the sample size representing the number of estimated propensity scores from each exposure group that must lie sufficiently close, that is, within an interval of length a to any given point in order for the point to be added to the RO. Framing this definition in a way that can be operationalized, it says that there is sufficient data support (overlap) at a point o if, for each exposure group separately, we can form a set that includes (1) o and (2) more than b estimated propensity scores and lies entirely within an interval of size less than a, that is, has range less than a.

We now introduce notation that will be used in the definition. Let N_e denote the number of units in exposure group e and $\hat{\xi}^e_{(i)}$ denote the ith propensity score order statistic in exposure group e. Using this notation, we propose the following definition for the region of overlap:

$$O = \{o \in P : \text{for } e = 0, 1, \text{range}(\{o, \hat{\xi}^e_{(i)}, \ldots, \hat{\xi}^e_{(i+b)}\}) < a \text{ for some } i = 1, \ldots, N_e - b\}.$$

This formalizes the notion introduced above of finding a set of o and more than b propensity scores, $\{o, \hat{\xi}^e_{(i)}, \ldots, \hat{\xi}^e_{(i+b)}\}$, with range less than a for each exposure group.

In comparison with previous definitions, our overlap definition provides a flexible, transparent and data-driven approach to identifying regions of poor data support with relatively easy-to-understand tuning parameters that give users the ability to decide what constitutes "sufficient" data support in the context of their own methods and application. Another contribution of this overlap definition to the wider literature is that it allows for regions of non-overlap in the interior of the propensity score distribution. In Figure 1 in Section 3 of the Supplementary Material (Nethery, Mealli and Dominici (2019)), we provide an illustration of the types of non-overlap that can be captured by this definition.

We define the RO and RN for BART + SPL using this definition. However, we note that BART + SPL is designed to handle exclusively non-overlap in the tails of the propensity score distribution. In Section 3.3 we provide guidance on how to specify a and b when applying BART + SPL.

2.2. *BART for causal inference.* BART (Chipman, George and McCulloch (2010)) is a Bayesian tree ensemble method that has been shown to have strong predictive performance in a variety of contexts (Bonato et al. (2011), Kindo, Wang and Peña (2016), Liu et al. (2015), Sparapani et al. (2016)). It is highly regarded for its consistently strong performance under the "default" model specifications reducing its dependence on subjective tuning and time consuming cross validation procedures. Letting j index the J trees in the ensemble ($j = 1, \ldots, J$), a BART is a sum of trees model of the form

$$Y = \sum_{j=1}^{J} g(\boldsymbol{X}; \mathcal{T}_j, \mathcal{M}_j) + \epsilon, \tag{1}$$

where g is a function that sorts each unit into one of a set of m_j terminal nodes, associated with mean parameters $\mathcal{M}_j = \{\mu_1, \ldots, \mu_{m_j}\}$ and based on a set of decision rules, $\mathcal{T}_j$. ϵ is a random error term that is typically assumed to be $N(0, \sigma^2)$ when the outcome is continuous. BART has also been extended to the binary outcome setting through the addition of the probit link function.

BART was introduced as a tool for causal inference by Hill (2011), who suggested using it to predict missing potential outcomes. Hahn et al. (2018) additionally recommend including the estimated propensity score as a covariate. Despite the accuracy of BART's potential outcome prediction in regions with strong data support, its predictions sometimes contain greater bias than those of parametric and classic causal inference methods in the presence of non-overlap (Hill (2011), Hill and Su (2013)). Because BART relies on binary cuts of the observed predictors, it is unable to capture trends in the data and therefore extrapolates poorly.

2.3. *BART + SPL.* In this section we describe BART + SPL, our proposed Bayesian approach for estimating causal effects in the presence of propensity score non-overlap. The first stage of the procedure, which we call the imputation phase,

utilizes a BART to impute the missing potential outcomes and estimate individual causal effects in the RO. In the second stage, which we call the smoothing stage, a spline is fit to the BART-estimated individual causal effects in the RO and is invoked to extrapolate the causal effect trends to the individuals in the RN leveraging the information from the observed potential outcomes for observations in the RN. Our approach for continuous outcomes is described in Sections 2.3.1 and 2.3.2 in the context of a single iteration of a Bayesian MCMC sampler for the sake of clarity, and in Section 2.3.3 we explain how the draws from the sampler can be invoked to estimate causal effects and uncertainties. This model can be implemented using the MCMC procedure described in Section 1 of the Supplementary Material (Nethery, Mealli and Dominici (2019)). Section 2.4 provides an extension to binary outcomes.

2.3.1. *Imputation stage*. In the first stage of BART + SPL, we adopt the common practice of treating the unobserved potential outcome for each individual as missing data, and we construct a BART model to impute these missing values for individuals in the RO. We introduce the subscripts q and r to index data from subjects in the RO and RN respectively, for example, Y_q^{obs} is the observed outcome of individual q in the RO, and Y_r^{mis} is the missing potential outcome of individual r in the RN ($q = 1, \ldots, Q; r = 1, \ldots, R$). Subscript O and subscript $O^\perp$ refer to vectors/matrices of the values of all individuals in O and $O^\perp$ respectively, for example, $\mathbf{Y}_O^{\text{mis}} = [Y_1^{\text{mis}}, \ldots, Y_Q^{\text{mis}}]'$ and $\mathbf{Y}_{O^\perp}^{\text{mis}} = [Y_1^{\text{mis}}, \ldots, Y_R^{\text{mis}}]'$.

In this stage all of our modeling efforts are focused on the data in the RO. Y_q^{mis} is first imputed using a BART model of the form

$$Y_q^{\text{obs}} = \sum_{j=1}^{J} g(E_q, \hat{\xi}_q, \boldsymbol{X}_q; \mathcal{T}_j, \mathcal{M}_j) + \epsilon_q,$$

where $\epsilon_q \sim N(0, \sigma_B^2)$. To do so, the Bayesian backfitting algorithm of Chipman, George and McCulloch (2010) is utilized to collect a sample from the posterior distribution of $\theta = \{\sigma_B^2, \mathcal{T}_j, \mathcal{M}_j; j = 1, \ldots, J\}$, $p(\theta|\mathbf{Y}_O^{\text{obs}})$. An imputed value of $\mathbf{Y}_O^{\text{mis}}$, denoted $\tilde{\mathbf{Y}}_O^{\text{mis}}$, is obtained by sampling from its posterior predictive distribution (ppd), $p(\mathbf{Y}_O^{\text{mis}}|\mathbf{Y}_O^{\text{obs}}) = \int p(\mathbf{Y}_O^{\text{mis}}|\mathbf{Y}_O^{\text{obs}}, \theta) p(\theta|\mathbf{Y}_O^{\text{obs}})\, d\theta$. $\mathbf{Y}_O^{\text{obs}}$ and $\tilde{\mathbf{Y}}_O^{\text{mis}}$ are used to form a sample of the individual causal effects in O, $\tilde{\boldsymbol{\Delta}}_O$.

2.3.2. *Smoothing stage*. In the second stage a smoothing model is fit to the BART-estimated individual causal effects in the RO, and the model is employed to estimate the individual causal effects in the RN by extrapolating the trends identified in the RO. With this approach we impose the assumption that any trends in the individual causal effects (as a function of the propensity score and/or the covariates) identified in the RO can be extended into the RN. By modeling the causal effect surface in this stage rather than the separate potential outcome surfaces, we

take advantage of the potentially increased smoothness of the causal effects that may occur in practice. Through tuning we ensure that the variance in the RN is inflated to reflect the high uncertainty in the region.

Assume for now that the RN includes only exposed individuals so that $E_r = 1$ for all r. Define $Y_q^*(1) = Y_q^{\text{obs}}$ if $E_q = 1$ and $Y_q^*(1) = \tilde{Y}_q^{\text{mis}}$ otherwise (remember $\tilde{Y}_q^{\text{mis}}$ denotes the BART-imputed missing potential outcome for q). Thus, $Y_q^*(1)$ is the observed or imputed potential outcome corresponding to $E = 1$ for each unit in the RO. Let $\text{rcs}(z)$ denote a restricted cubic spline basis for z. Employing the imputed values obtained in the previous stage, we construct the following smoothing stage model:

$$\tilde{\Delta}_q = \boldsymbol{W}'\boldsymbol{\beta} + \epsilon_q, \qquad \boldsymbol{W} = \begin{bmatrix} \text{rcs}(\hat{\xi}_q) \\ \text{rcs}(Y_q^*(1)) \\ \boldsymbol{X}_q \end{bmatrix},$$

where $\epsilon_q \sim N(0, \sigma_S^2 + I(\hat{\xi}_q \in O^\perp)\tau_q)$. σ_S^2 is the residual variance for all units in the RO and τ_q is an added variance component only applied to units in the RN. The purpose of τ_q is to inflate the variance of units with an estimated propensity score in the RN to adequately reflect the higher uncertainty in regions of little data support. In this model τ_q is clearly unidentifiable, as the model is fit using exclusively data in the RO, and it will only come into play when invoking the ppd to predict in the RN. Here, we choose to treat it as a tuning parameter, and below we describe our recommended tuning parameter specification. We recommend restricted cubic splines in the smoothing model, because they generally demonstrated superior performance when applied to simulated data. We also recommend excluding a small portion of the data at the tails of the RO (i.e., tails of the propensity score in the RO) from spline model fitting, either through the use of boundary knots or by omitting these data from the model, because BART's predictions in the tails of variables can be unstable and can negatively affect the SPL's performance.

We collect a posterior sample of the spline parameters, $\psi = \{\boldsymbol{\beta}, \sigma_S^2\}$. Recall the motivation for the smoothing stage is to use the trends from the RO to predict the individual causal effects in the RN. Thus, the sampled ψ and the covariate values for individuals in the RN are summoned to obtain a sample from the ppd of $\boldsymbol{\Delta}_{O^\perp}$, $p(\boldsymbol{\Delta}_{O^\perp}|\boldsymbol{\Delta}_O) = \int p(\boldsymbol{\Delta}_{O^\perp}|\boldsymbol{\Delta}_O, \psi)p(\psi|\boldsymbol{\Delta}_O)\,d\psi$. The sample is denoted $\tilde{\boldsymbol{\Delta}}_{O^\perp}$.

Note that including $Y_q^*(1)$ as a predictor permits the model to capture the relationship between the causal effects and $Y(1)$, the potential outcome observed for all units in the RN. This allows the observed potential outcomes in the RN to aid in the extrapolation, so that this information is not wasted. In the case that both exposed and unexposed units fall in the RN, we define $Y_q^*(0)$ analogously to $Y_q^*(1)$ and construct a second model, identical to the one above, except replacing $\text{rcs}(Y_q^*(1))$ with $\text{rcs}(Y_q^*(0))$. The ppd from the first model is then used to predict individual causal effects for exposed units in the RN, and the ppd from the second model is used for unexposed units.

Our recommended specification of τ_q is motivated by the aim to have (1) the variance of individual causal effects increase monotonically as the observation's distance from the RO (i.e., region of strong data support) increases and (2) the increase in variance be in proportion to the scale of the data. Thus, the suggested tuning parameter specification is $\tau_q = (10d_q)t_O$, where $t_O = \text{range}(\tilde{\mathbf{\Delta}}_O)$ and d_q is the distance from the observation's propensity score to the nearest propensity score in the RO. The effect of this tuning parameter is that, for every .1 unit further we go into the RN, the variance of the individual causal effects increases by the range of the causal effects in the RO. While this may produce conservative uncertainties in situations where the trends in the RN are easily predicted from the trends in the RO, we have found in simulations that this choice of tuning parameter consistently provides both reasonably-sized credible intervals and acceptable coverage.

2.3.3. *Estimation and uncertainty quantification*. We can iterate the two stages described above M times to obtain $\{\mathbf{\Delta}_O^{(1)}, \ldots, \mathbf{\Delta}_O^{(M)}\}$ from the imputation stage and $\{\mathbf{\Delta}_{O^\perp}^{(1)}, \ldots, \mathbf{\Delta}_{O^\perp}^{(M)}\}$ from the smoothing stage (note that we have traded the tilde notation from above for the (m) notation to differentiate the samples from the M iterations). By iterating between the two stages we are able to account for the uncertainty in the estimation of $\mathbf{\Delta}_O$ from the first stage and pass it on to the second stage, where $\mathbf{\Delta}_O$ is used as the outcome. Thus, the uncertainty in the estimate of $\mathbf{\Delta}_{O^\perp}$ reflects the uncertainty both from stage one and stage two.

For units in the RO and the RN, individual causal effects are estimated as $\hat{\Delta}_q = \frac{1}{M}\sum_{m=1}^{M} \Delta_q^{(m)}$ and $\hat{\Delta}_r = \frac{1}{M}\sum_{m=1}^{M} \Delta_r^{(m)}$ respectively, that is, the posterior mean over the M samples. Credible intervals for the individual causal effects can be obtained by extracting the appropriate percentiles from these M samples. Samples of Δ_S are produced by $\Delta_S^{(m)} = \frac{1}{N}(\sum_{q=1}^{Q} \Delta_q^{(m)} + \sum_{r=1}^{R} \Delta_r^{(m)})$ for $m = 1, \ldots, M$, and $\hat{\Delta}_S = \frac{1}{M}\sum_{m=1}^{M} \Delta_S^{(m)}$. As above, percentiles of the M samples provide credible interval for Δ_S.

In order to estimate Δ_P, an additional integration over the predictors is required. Wang et al. (2015) discuss the necessity of such an integration step when estimating population average causal effects with models that permit nonlinearity and/or heterogeneity, and they propose the application of the Bayesian bootstrap to execute it. We adopt the same approach here. For each sample of the individual causal effects, $\{\mathbf{\Delta}_O^{(m)}, \mathbf{\Delta}_{O^\perp}^{(m)}\}$, the Bayesian bootstrap is performed on it B times (where B is a large constant) and the average of each bootstrap sample taken to obtain B draws from the posterior distribution of the population average causal effect. We randomly select one of these samples and call it $\Delta_P^{(m)}$, so that in the end we have collected $\{\Delta_P^{(1)}, \ldots, \Delta_P^{(M)}\}$. The population average causal effect is estimated as $\hat{\Delta}_P = \frac{1}{M}\sum_{m=1}^{M} \Delta_P^{(m)}$ and the credible interval formed using percentiles.

2.4. *BART + SPL with binary outcomes.* By invoking BART probit (Chipman, George and McCulloch (2010)) in the imputation stage and utilizing a simple arcsine transformation in the smoothing stage, we can straightforwardly extend BART + SPL to the binary outcomes setting. While most of our notation will remain the same for binary outcomes, we note a few changes. Individual causal effects are traditionally defined as the difference in each individual's potential outcomes, as above; however, in the binary outcomes setting, estimating those differences, which can only take on values -1, 0 and 1, may be challenging and may sacrifice information. Because, in our approach, we are treating the potential outcomes as random variables, it is reasonable and desirable to instead define the individual causal effects as differences in some features of the distributions of the potential outcomes, although doing so requires a slight abuse of traditional terminology/notation. For binary outcomes we define the individual causal effects as $\Delta_i = P(Y_i(1) = 1) - P(Y_i(0) = 1)$ and the estimands as $\Delta_S = \frac{1}{N}\sum_{i=1}^{N} \Delta_i$, $\Delta_{P|\boldsymbol{x}} = P(Y(1) = 1|\boldsymbol{x}) - P(Y(0) = 1|\boldsymbol{x})$ and $\Delta_P = E_{\boldsymbol{X}}[P(Y(1) = 1|\boldsymbol{X}) - P(Y(0) = 1|\boldsymbol{X})]$. Here, we fit a BART probit to estimate individual causal effects in O, fit a spline model to the arcsine transform of these estimates (which are bounded between -1 and 1) and use the spline to estimate individual causal effects for units in $O^{\perp}$. While we provide below explicit forms for the imputation and smoothing models in the binary setting, we refer the reader back to the previous section for the full sampling procedure details which follow analogously to the continuous outcomes setting.

In the imputation stage the BART probit model fit to the RO data has the following form:

$$P(Y_q^{\text{obs}} = 1) = \Phi\left(\sum_{j=1}^{J} g(E_q, \hat{\xi}_q, \boldsymbol{X}_q; \mathcal{T}_j, \mathcal{M}_j)\right),$$

where $\Phi()$ is the standard normal cumulative distribution function. With this model posterior samples $\tilde{P}(Y_q^{\text{obs}} = 1)$ and $\tilde{P}(Y_q^{\text{mis}} = 1)$ can be drawn and used to form a posterior sample of the individual causal effect, $\tilde{\Delta}_q$.

For the smoothing stage, as above, assume without loss of generality that all the units in $O^{\perp}$ all have $E_r = 1$. Define $Y_q^*(1) = Y_q^{\text{obs}}$ if $E_q = 1$ and $Y_q^*(1) = I(\tilde{P}(Y_q^{\text{mis}} = 1) > 0.5)$ otherwise. Then the smoothing model is

$$\text{arcsine}(\tilde{\Delta}_q) = \boldsymbol{W}'\boldsymbol{\beta} + \epsilon_q, \qquad \boldsymbol{W} = \begin{bmatrix} \text{rcs}(\hat{\xi}_q) \\ Y_q^*(1) \\ \boldsymbol{X}_q \end{bmatrix},$$

where $\epsilon_q \sim N(0, \sigma_S^2)$. Note that, unlike in the continuous case, no tuning parameter is included in the variance, as simulations indicated it was not needed to obtain reasonable coverage in the binary setting. Individual causal effects on the arcsine scale for units in $O^{\perp}$ can be obtained from the posterior predictive distribution

and back-transformed to the desired scale. As described in the previous section, a second analogous smoothing model can be fit if $O^{\perp}$ contains units from both the exposed and unexposed groups. Average causal effect estimation and uncertainty quantification proceed identically to the continuous case.

3. Simulations. In this section we conduct simulation studies to evaluate the performance of BART + SPL relative to existing methods in the presence of non-overlap and to provide guidance on how to specify parameters a and b in the non-overlap definition to obtain optimal performance. In Section 3.1 we simulate data with a small number of confounders and varying degrees of non-overlap, and we compare BART + SPL's population average causal effect estimation performance to that of a standard BART and of an existing spline-based method for causal inference. In Section 3.2 we generate data with non-overlap and high dimensional covariates and compare the performance of BART + SPL and the spline-based method. Finally, data with varying amounts of non-overlap are generated and BART + SPL is implemented with various specifications of a and b to provide insight on the optimal choices in Section 3.3. R code (R Core Team (2016)) to implement BART + SPL and to reproduce all simulations is available on Github at https://github.com/rachelnethery/overlap.

3.1. *Performance of BART + SPL relative to existing methods.* We purposely simulate data under a challenging situation of: a) propensity score non-overlap; b) nonlinearity of the potential outcomes in the propensity score; and c) heterogeneous causal effects. We wish to evaluate the relative performance of our method when utilizing a true propensity score and when utilizing a misspecified propensity score estimate. We first discuss the simulation structure when utilizing the true propensity score. We let $N = 500$ and assign half of the subjects to $E = 1$. We generate two confounders that are highly associated with the exposure (E), one binary ($X_1 : X_1|E = 1 \sim \text{Bernoulli}(0.5), X_1|E = 0 \sim \text{Bernoulli}(0.4)$) and one continuous ($X_2 : X_2|E = 1 \sim N(2 + c, \sqrt{1.25 + 0.1c}), X_2|E = 0 \sim N(1, 1)$). Given these specifications, the true propensity scores can easily be calculated using Bayes Rule. The potential outcomes are constructed as $Y_i(1) = -3(1 + \exp(-(10(X_{2i} - 1)))^{-1} + 0.25X_{1i} - X_{1i}X_{2i}$ and $Y_i(0) = -1.5X_{2i}$. We label the simulations with the true propensity score as 3.1A.

For the simulations with a misspecified propensity score estimate, we again let $N = 500$ and assign half of the subjects to $E = 1$. We generate a binary confounder ($X_1 : X_1|E = 1 \sim \text{Bernoulli}(0.5), X_1|E = 0 \sim \text{Bernoulli}(0.4)$) and a continuous confounder ($X_2 : X_2|E = 1 \sim N(2 + c, 4), X_2|E = 0 \sim N(1, 1)$). The potential outcomes are $Y_i(1) = 3(1 + \exp(-(10(X_{2i} - 1)))^{-1} + 0.25X_{1i} - 0.1X_{1i}X_{2i} + 0.5$ and $Y_i(0) = 0.2X_{2i} + 0.1X_{2i}^2 + 1$. As is common in the literature, we use a simple logistic regression model of the form $\text{logit}(P(E_i = 1)) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$ to estimate the propensity scores. This model is clearly misspecified, because, for ex-

ample, the true relationship between E and X_2 is not linear. In practice when the form of the propensity score model is unknown, we encourage the use of flexible models for estimation (Westreich, Lessler and Funk (2010)), such as BART, neural networks or support vector machines in order to reduce the chance of propensity score model misspecification. The use of BART for propensity score estimation is demonstrated in the application to real data in Section 4. Various flexible propensity score estimation methods could be tested and the method that achieves the best covariate balance selected. We label the simulations with the misspecified propensity score as 3.1B.

We have selected these simulation structures so that, for a single value of c, the type and degree of propensity score non-overlap in 3.1A and 3.1B should be similar. Both 3.1A and 3.1B produce data sets with lack of overlap in the right tail of the propensity score distribution (i.e., individuals from the unexposed group are unobserved or very sparse) and with varying degrees of non-overlap controlled by c. Our simulations are designed to produce non-overlap in the right tail of the propensity score distribution and our motivation is to demonstrate how our method performs in the presence of different features in this RN. Thus, simulated datasets are utilized in the results below only if any intervals of non-overlap outside the right tail contain 10 observations or fewer (cumulatively), and, in these datasets, the intervals of non-overlap outside the right tail are ignored (i.e., treated as part of the RO). In this way we ensure that the results solely reflect how the tested methods respond to the features of the intended RN.

We consider three separate simulated scenarios, that is, three different specifications of c, within 3.1A and 3.1B. We let $c = 0$ (simulations 3.1A-i and 3.1B-i), $c = 0.35$ (simulations 3.1A-ii and 3.1B-ii) and $c = 0.7$ (simulations 3.1A-iii and 3.1B-iii). Example datasets from each are illustrated in Figure 2 in Section 3 of the Supplementary Material (Nethery, Mealli and Dominici (2019)). With $c = 0$, the RN is quite small, and the trend in the individual causal effects in the RN is mildly nonlinear. With $c = 0.35$, the RN is somewhat larger, and the trends exhibited by the individual causal effects in the RN are moderately nonlinear. With $c = 0.7$, a substantial portion of the sample lies in the RN, and the causal effects in the RN are highly nonlinear. We use our definition of overlap with $a = 0.1$ and $b = 7$ to define the RO and RN for each simulated dataset.

We implement BART + SPL on 1000 simulated datasets under each condition. Gutman and Rubin (2015) recommended a spline-based multiple imputation approach for estimating average causal effects. We compare the performance of BART + SPL versus Gutman and Rubin's method with and without trimming (T-GR and U-GR respectively) and also BART with and without trimming (T-BART and U-BART respectively). Detailed results of the untrimmed analyses appear in Table 1 and the distributions of the average causal effect estimates from the trimmed and untrimmed analyses can be compared in Figure 3 in Section 3 of the Supplementary Material (Nethery, Mealli and Dominici (2019)).

TABLE 1
Absolute (Abs) bias, 95% credible interval coverage and mean square error (MSE) in estimation of the population average causal effects in simulations from Section 3.1

Simulation setting	Method	Abs Bias (%)	Coverage	MSE
3.1A-i	U-GR	0.12 (46.46)	0.33	1.25
	U-BART	0.03 (10.72)	0.99	0.07
	BART + SPL	0.01 (5.63)	1.00	0.05
3.1A-ii	U-GR	0.17 (97.31)	0.23	1.69
	U-BART	0.05 (31.95)	0.89	0.12
	BART + SPL	0.02 (12.58)	1.00	0.09
3.1A-iii	U-GR	0.23 (724.78)	0.19	2.34
	U-BART	0.08 (427.55)	0.71	0.22
	BART + SPL	0.03 (100.80)	1.00	0.14
3.1B-i	U-GR	0.26 (50.42)	0.00	0.64
	U-BART	0.13 (25.47)	0.12	0.33
	BART + SPL	0.11 (21.64)	0.62	0.27
3.1B-ii	U-GR	0.32 (66.58)	0.00	0.73
	U-BART	0.18 (36.73)	0.04	0.48
	BART + SPL	0.15 (30.95)	0.55	0.36
3.1B-iii	U-GR	0.41 (94.79)	0.00	0.89
	U-BART	0.24 (55.00)	0.01	0.68
	BART + SPL	0.21 (48.03)	0.35	0.49

The simulation results demonstrate the dominant performance of BART + SPL compared to U-BART and U-GR under a wide range of challenging conditions. However, in extreme scenarios with unpredictable trends and large portions of the sample in the RN, even the performance of BART + SPL may deteriorate, as demonstrated by simulation 3.1A-iii, where BART + SPL gives high percent bias, and in simulation 3.1B-iii, where BART + SPL gives high percent bias and poor coverage. Nonetheless, BART + SPL's performance still exceeds that of its competitors. For both BART and GR the trimmed estimates, which are no longer estimators of the population level causal effects, are further from the true population average causal effects than the untrimmed estimates.

We also conducted a simulation study to evaluate the performance of BART + SPL for binary outcomes. The data and results are described in Section 2 of the Supplementary Material (Nethery, Mealli and Dominici (2019)). BART + SPL performed similar to or better than the competing methods (U-GR and U-BART) in each of our simulations with $N = 500$. However, even without non-overlap, BART probit can fail to provide improvements over parametric methods when sample sizes are small to moderate, and thus we recommend that BART + SPL for binary outcomes only be applied to large datasets (i.e., $N \geq 500$).

3.2. *BART + SPL with high dimensional covariates*. One of the most widely-recognized limitations of BART is its poor performance when the number of predictors, p, is large (Chipman, George and McCulloch (2010)). The decline in performance is most significant when many irrelevant predictors (i.e., predictors unrelated to the outcome) are included. Thus, in this section we seek to examine whether and how BART + SPL should be applied in settings where the number of potential confounders is large. Although BART has been extended to permit sparsity in the $p > N$ setting (Linero (2018)), we do not consider the $p > N$ case here.

For these simulations we let $N = 500$ and assign half of the sample to $E = 1$. We then generate 10 confounders, five binary and five continuous. The binary confounders have distribution $X_1|E = 1, \ldots, X_5|E = 1 \sim \text{Bernoulli}(0.45)$, $X_1|E = 0, \ldots, X_5|E = 0 \sim \text{Bernoulli}(0.4)$, and the continuous confounders have distribution $X_6|E = 1, \ldots, X_{10}|E = 1 \sim N(2, 4)$, $X_6|E = 0, \ldots, X_{10}|E = 0 \sim N(1.3, 1)$. We consider the following three scenarios: only these 10 confounders are present (simulation 3.2A); these 10 confounders as well as 25 randomly generated "potential confounders" are present (simulation 3.2B); and these 10 confounders as well as 50 randomly generated "potential confounders" are present (simulation 3.2C). Of course in real applications, we often do not know a priori which of the potential confounders are true confounders; hence, we include them all in the modeling. A propensity score is formed using predicted probabilities from the logistic regression $\text{logit}(P(E_i = 1)) = \beta_0 + \boldsymbol{Z}_i\boldsymbol{\beta}$, where $\boldsymbol{Z}_i$ is a vector of the true and potential confounders. The potential outcomes are generated so that they exhibit nonlinear trends in the estimated propensity score—$Y_i(0) = 0.5(X_{1i} + X_{2i} + X_{3i} + X_{4i} + X_{5i}) + 15(1 + \exp(-8X_{6i} + 1))^{-1} + X_{7i} + X_{8i} + X_{9i} + X_{10i} - 5$ and $Y_i(1) = X_{1i} + X_{2i} + X_{3i} + X_{4i} + X_{5i} - 0.5(X_{6i} + X_{7i} + X_{8i} + X_{9i} + X_{10i})$.

The features of these data are illustrated in Figure 5 in Section 3 of the Supplementary Material (Nethery, Mealli and Dominici (2019)). The simulations are designed to have a large RN in the right tail of the estimated propensity score with moderate nonlinearity in the causal effect in the RN. The RO and RN are defined using tuning parameters $a = 0.1$ and $b = 7$.

We simulate 1000 datasets from each of the three scenarios described above. We apply both BART + SPL and the untrimmed Gutman and Rubin spline method (GR) to each dataset. Results are provided in Table 2.

These results reflect BART's struggle in the presence of irrelevant predictors. When only the 10 true confounders are included in the modeling, BART + SPL outperforms GR and demonstrates similar performance as in Section 3.1. However, when irrelevant predictors are introduced, GR's bias decreases while BART + SPL's remains constant or increases. With 50 irrelevant predictors GR's bias is substantially lower than BART + SPL's (although, notably, its coverage and MSE remain inferior). This suggests that BART + SPL is only likely to improve on existing methods in settings where the set of true confounders can be posited a priori with some confidence.

TABLE 2
Absolute (Abs) bias, 95% credible interval coverage and mean square error (MSE) in estimation of the population average causal effects in simulations from Section 3.2

Simulation setting	Method	Abs Bias (%)	Coverage	MSE
3.2A	GR	0.98 (5.64)	0.03	13.14
	BART + SPL	0.56 (3.22)	0.99	8.97
3.2B	GR	0.56 (3.24)	0.40	14.32
	BART + SPL	0.56 (3.20)	0.97	8.49
3.2C	GR	0.36 (2.10)	0.73	16.09
	BART + SPL	0.64 (3.66)	0.96	9.69

3.3. *Guidelines for defining the RN.* The simulation results in this section are intended to provide guidance on both the degree of non-overlap that threatens BART's performance and the degree of non-overlap that threatens BART + SPL's performance. They also suggest appropriate default specifications of tuning parameters a and b in the non-overlap definition. To impose strict control on the size of the RO and RN, in these simulations we utilize a single confounder rather than a propensity score. Based on the above simulations, we expect the performance of BART + SPL to be comparable with varying numbers of confounders, as long as few irrelevant covariates are included.

We let $N = 500$ and assign half of the sample to $E = 1$. We generate the confounder as $X|E = 1 \sim N(2.5, 4)$, $X|E = 0 \sim N(v, w)$, where v and w control the degree of non-overlap. Unlike in the previous simulations, in these we generate a single, fixed instance of the confounder and simply add random noise to (a function of) it to create the potential outcomes for each simulation. We consider two potential outcome scenarios, one of which produces data that are relatively simple to model (with BART) while the other produces data that are challenging to model. The former, which we label simulation 3.3A, is created by assigning $Y(0) = 1.5 + \frac{X+(X^2/2!)}{20} + \mathrm{N}(0, 0.06)$ and $Y(1) = \frac{1}{1+e^{-(X-1)}} + \mathrm{N}(0, 0.06)$ and the latter, which we call simulation 3.3B, by $Y(0) = 1.5 + \frac{X+(X^2/2!)+(X^3/3!)}{20} + \mathrm{N}(0, 0.06)$ and $Y(1) = \frac{1}{1+e^{-(X-1)}} + \mathrm{N}(0, 0.06)$.

In both simulation 3.3A and 3.3B we achieve different degrees of non-overlap, primarily non-overlap in the right tail of the confounder, by manipulating v and w. In order from least to most non-overlap, we consider $\{v = 1.4, w = 1.96\}$, $\{v = 0.75, w = 1.44\}$ and $\{v = 0, w = 1\}$. Moreover, in each scenario we test the following three specifications of $\{a, b\}$ in the overlap definition in order from most to least conservative: $\{a = 0.05 * (\mathrm{range}(X)), b = 10\}$, $\{a = 0.1 * (\mathrm{range}(X)), b = 10\}$, and $\{a = 0.15 * (\mathrm{range}(X)), b = 3\}$. Each combination of $\{v, w\}$ and $\{a, b\}$ leads to nine different settings for each of simulation 3.3A and 3.3B for a total of 18 simulation settings. The percent of the sample falling into the RN, denoted π

TABLE 3

Absolute (Abs) bias, 95% credible interval coverage (Covg), and mean square error (MSE) in estimation of the population average causal effects using BART + SPL and BART applied to simulations 3.3A and 3.3B. BART + SPL-1 refers to BART + SPL with the RO defined as $a = 0.05 * (\text{range}(X)), b = 10$, BART + SPL-2 refers to BART + SPL with the RO defined as $a = 0.1 * (\text{range}(X)), b = 10$, and BART + SPL-3 refers to BART + SPL with the RO defined as $a = 0.15 * (\text{range}(X)), b = 3$

		Simulation 3.3A			Simulation 3.3B		
v, w	Method (π)	Abs Bias (%)	Covg	MSE	Abs Bias (%)	Covg	MSE
1.4, 1.96	BART + SPL-1 (17)	0.03 (2.51)	1.00	0.08	0.05 (4.49)	1.00	0.12
	BART + SPL-2 (8)	0.03 (2.23)	1.00	0.07	0.04 (4.15)	1.00	0.11
	BART + SPL-3 (2)	0.03 (2.43)	1.00	0.08	0.04 (3.97)	1.00	0.10
	BART	0.03 (2.98)	0.81	0.09	0.05 (4.69)	0.57	0.12
0.75, 1.44	BART + SPL-1 (24)	0.05 (4.14)	1.00	0.09	0.07 (6.27)	1.00	0.14
	BART + SPL-2 (14)	0.04 (3.44)	1.00	0.08	0.06 (5.72)	1.00	0.13
	BART + SPL-3 (6)	0.05 (4.06)	1.00	0.09	0.06 (5.83)	1.00	0.12
	BART	0.06 (5.50)	0.47	0.11	0.07 (6.71)	0.35	0.14
0, 1	BART + SPL-1 (34)	0.08 (6.65)	1.00	0.11	0.09 (8.27)	1.00	0.16
	BART + SPL-2 (21)	0.05 (4.60)	1.00	0.09	0.06 (5.72)	1.00	0.12
	BART + SPL-3 (11)	0.06 (4.71)	1.00	0.09	0.06 (5.12)	1.00	0.11
	BART	0.07 (6.28)	0.59	0.11	0.06 (5.67)	0.74	0.12

in these simulations ranges from $\pi = 2\%$ to $\pi = 34\%$. Of course the impact of non-overlap on average causal effect estimates depends not only on the proportion of the sample falling in the RN but also likely on the extremity of the observations in the RN relative to the RO. In our simulations, as π increases, the average distance between observations in the RO and the RN also increases. An example dataset from both simulation 3.3A and simulation 3.3B is presented in Figure 6 in Section 3 of the Supplementary Material (Nethery, Mealli and Dominici (2019)).

We apply a standard BART (ignoring the non-overlap) and BART + SPL to 1000 simulated datasets under each of the 18 conditions. Table 3 contains the results for simulations 3.3A and 3.3B. While BART + SPL nearly always performs better in terms of each metric than BART, the most notable difference in the BART + SPL and BART results is the difference in coverage probabilities, with BART + SPL consistently obtaining conservative coverage and BART's coverage deteriorating as the degree of non-overlap increases. Even when only 2% of the data falls into the RN, BART's coverage is unreliable. Thus, BART could provide misleading inference even with small amounts of non-overlap.

BART + SPL's coverage is conservative, but reliable, in all the simulations assessed; however, its bias tends to increase as the degree of non-overlap increases. Thus, it appears that, if some bias in the point estimate can be tolerated, BART + SPL can be expected to provide conservative inference in (nonpatho-

logical) scenarios with over 25% of the data in the RO. However, based on the observation that BART + SPL's bias is greater than 5% in simulation 3.3B under each overlap definition with $\{v = 0.75, w = 1.44\}$ and $\{v = 0, w = 1\}$, a more conservative option might be to sacrifice the population-level estimand and performed a trimmed or weighted analysis when more than 15% of the data falls in the RN.

Finally, these simulations suggest that BART + SPL is quite robust to the specification of a and b, as we see relatively small discrepancies in bias and coverage. However, some of the results of 3.3B indicate that one should avoid defining the RO too conservatively, as discarding too much information can lead to modest increases in bias. The moderate choice of $a = 0.1 * \text{range}(X)$ and $b = 10$ provides the best results in most of the simulations, and we, therefore, recommend this as the default specification.

4. The effect of natural gas compressor stations on county-level thyroid cancer and leukemia mortality. We collected 2014 thyroid cancer and leukemia mortality rate estimates for each county in the US from the Global Health Data Exchange. The data and methods used to develop these estimates have been described previously (Mokdad et al. (2017)). We also obtained the locations of NG compressor stations from publicly available data complied by Oak Ridge National Laboratory (2017). While the data is not guaranteed to be complete, it is, to our knowledge, the most comprehensive documentation of compressor station locations in existence with 1359 compressor station locations verified using imagery. In order to test a causal hypothesis, we need to assume that exposure to compressor station-related emissions preceded 2014 (the year for which cancer mortality rates are observed) by at least the minimum latency period for thyroid cancer and leukemia. The CDC reports the minimum latency period for thyroid cancer as 2.5 years and the minimum latency for leukemia as 0.4 years (World Trade Center Health Program (2015)). Although the dataset does not contain dates of origin for the compressor stations, it does contain peak operation dates. Eighty-four percent of the compressor stations in the dataset have peak operating dates in or before 2012; thus, it seems reasonable to assume that most of the compressor stations in the dataset operated at least 2.5 years prior to 2014.

Our county-level exposure variable is an indicator of whether a compressor station is present in the county. We collected county-level demographic, socioeconomic and behavioral confounder data from the American Community Survey 2014 five-year estimates (US Census Bureau (2014)) and the 2014 County Health Rankings and Roadmaps (Robert Wood Johnson Foundation (2014)). Data were accessed using Social Explorer. The confounders used are rate of primary care physicians, percent of less than 65-year-olds uninsured, percent diabetic, percent current smokers, percent of people with limited access to healthy foods, percent obese, food environment index, population density, percent male, percent less than age 55, percent white, average household size, percent with bachelor's degree or higher, percent unemployed, median household income, Gini index of inequality, percent owner-occupied housing units, median rent as proportion of income and

average commute time to work. All the data used in this analysis are publicly available, and the data and R code to reproduce the analysis are posted on Github at https://github.com/rachelnethery/overlap.

We note that the sensitivity of this analysis to detect exposure effects will be low, because any true health effects of exposure to compression station emissions is likely more spatially concentrated than the county level. Although a higher spatial resolution analysis would be preferable, obtaining important behavioral confounder data at a finer spatial resolution is challenging. In an effort to improve the detectability of effects, we focus our analysis on roughly the mid-western region of the US (counties with centroid longitudes between -110 and -90), where few other sources of pollution exist compared to the coastal regions (Di et al. (2017b)). A focus on this region is also reasonable because NG production has a longer history in this region compared to other US production regions, likely leading to greater exposure.

We begin with a dataset of 1309 counties and, after discarding counties with any missing confounders, are left with $N = 978$ counties. Two hundred ninety-one of these counties are exposed (i.e., contain at least one compressor station) while 687 are unexposed. Table 2 in Section 3 of the Supplementary Material (Nethery, Mealli and Dominici (2019)) shows the differences in the exposed and unexposed populations. Notably, exposed counties have, on average, higher percent uninsured, lower population density, lower percent white, lower education and higher percent unemployment. We estimate a propensity score by applying a BART probit with exposure status as the response and all the confounders as predictors. The histogram in Figure 1 illustrates the non-overlap in the resulting

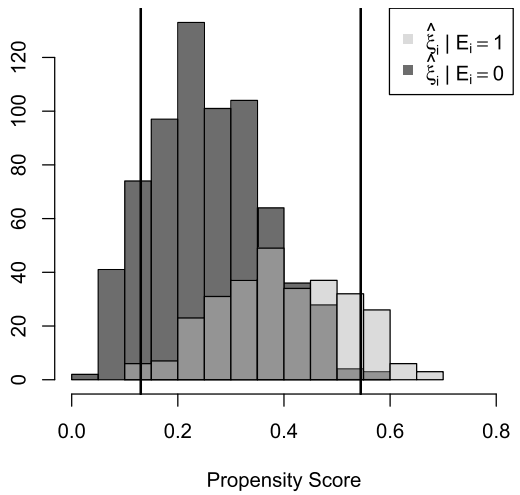

FIG. 1. *Estimated propensity score histograms stratified by exposure status and overlaid. Bold vertical lines represent the start of non-overlap intervals in both tails of the distribution.*

propensity score with the solid vertical lines denoting the start of the intervals of non-overlap that are detected in each tail of the propensity score using our overlap definition and $a = 0.1 * \text{range}(\hat{\xi})$ and $b = 10$. With these specifications 12% of the sample falls into the RN. BART + SPL is needed to obtain population level causal effect estimates in this setting.

The following outcome variables are considered: (1) 2014 thyroid cancer mortality rate; (2) the change in thyroid cancer mortality rate from 1980 to 2014; (3) 2014 leukemia mortality rate; and (4) the change in leukemia mortality rate from 1980 to 2014. The 2014 rates are log transformed prior to analysis. We analyze each outcome using both BART + SPL and trimmed BART. Counties in the trimmed sample are more urban and densely populated, on average, than the population represented by the full sample (trimmed sample average population density is 108.40 per mi^2 compared to 99.38 in the full sample).

Average causal effect estimates and 95% credible intervals from each analysis can be found in Table 4. The BART + SPL analysis is estimating population average causal effects, and the trimmed BART is estimating trimmed sample average causal effects. Two of the analyses find statistically significant effects of NG compressor stations—the trimmed BART analyses of the change in thyroid cancer mortality rates from 1980 to 2014 and of the change in leukemia mortality rates from 1980 to 2014. We must interpret these as significant effects only in the trimmed sample which is on average more urban than the population of interest. The population-level estimates from BART + SPL have wider credible intervals for two reasons. First, the additional marginalization over the confounders required to obtain population-level estimates increases the variance. Second, to estimate at the population level, we must account for the additional uncertainty induced by the

TABLE 4

Average causal effects of natural gas compressor station presence on 2014 *county-level thyroid cancer and leukemia mortality rates and the change in thyroid cancer and leukemia mortality rates from* 1980 *to* 2014

Outcome	Method	Effect	95% CI
2014 Thyroid Rates	BART + SPL	0.001	−0.017, 0.020
	BART	0.003	−0.007, 0.012
Change in Thyroid Rates 1980-2014	BART + SPL	0.992	−0.308, 2.237
	BART	1.089	0.130, 2.038
2014 Leukemia Rates	BART + SPL	0.006	−0.013, 0.025
	BART	0.005	−0.004, 0.014
Change in Leukemia Rates 1980-2014	BART + SPL	0.913	−0.361, 2.206
	BART	0.988	−0.014, 1.958

non-overlap, which BART + SPL does by inflating variances in the RN. With these wide credible intervals evidence of an effect must be very compelling in order to achieve statistical significance. However, the effect estimate from each analysis is positive, indicating a harmful effect of compressor stations, and the estimates are similar in the trimmed and untrimmed analyses.

In Tables 3 and 4 in Section 3 of the Supplementary Material (Nethery, Mealli and Dominici (2019)), we provide the results of two sensitivity analyses using alternate specifications of a and b, one resulting in a larger RN and one in a smaller RN. The BART + SPL results demonstrate little sensitivity to these choices with inference remaining the same in each analysis. This robustness agrees with findings in simulated data in Section 3.3. The trimmed BART, however, is sensitive to the a and b choices. The significant leukemia effect is attenuated in one sensitivity analysis, and both significant effects are attenuated in the other. This sensitivity is not surprising, given that changes in the observations trimmed correspond to changes in the estimand. Because the trimmed BART is sensitive to these choices, we should interpret the results with caution.

The significant and near-significant findings presented here suggest that the health effects of compressor station exposure is a topic that warrants further study with higher quality data. In particular an analysis at higher spatial resolution is needed to detect geographically concentrated effects that may be washed out at the county level. Moreover, counties with compressor stations may also be more likely to be located in NG production regions; thus, the effects of compressor station exposure may not be distinguishable from the effects of NG drilling and production-related exposures at the county level. Finally, an investigation of cancer diagnosis rates may be more informative than our study of mortality rates. However, cancer diagnosis rates are difficult to obtain across large geographic regions.

5. Discussion. In this paper we have introduced a general definition of propensity score non-overlap and have proposed a Bayesian modeling approach to estimate population average causal effects and corresponding uncertainties in the presence of non-overlap. A novel feature of our approach is its separation of the tasks of estimating causal effects in the region of overlap and the region of non-overlap. A nonparametric BART model is used to estimate individual causal effects in the region of overlap where there is strong data support. In the region of non-overlap, where reliance on model specification is required to estimate causal effects, individual causal effects are estimated by extrapolating trends from the region of overlap via a parametric spline model. BART + SPL can be applied to data with either continuous or binary outcomes and is implemented in a fully Bayesian manner, so that all sources of uncertainty are captured.

We demonstrated via simulations that BART + SPL outperforms both stand-alone BART and stand-alone spline causal inference approaches in estimation of population average causal effects under a wide range of conditions involving

propensity score non-overlap. However, due to BART's limitations in high dimensional settings, BART + SPL may give more biased results than existing methods when many irrelevant predictors are present.

While we have focused primarily on the use of our overlap definition with BART + SPL, it can also be used to define the RO for trimming, and it may provide a more transparent and reproducible approach than trimming by eye. However, as demonstrated by the trimmed BART results in Section 4, our overlap definition is unlikely to produce an interpretable trimmed estimand. Therefore, other strategies that prioritize the interpretability of the resulting estimand may be preferable (Li, Morgan and Zaslavsky (2018)).

We again note that BART + SPL is intended to handle finite sample non-overlap only (estimating population average causal effects may be inappropriate when the population exhibits non-overlap). When non-overlap is a finite sample feature, it will disappear asymptotically. For this reason we have avoided discussions of double robustness which is an asymptotic property. However, we have demonstrated in simulations that, relative to the competing methods considered, BART + SPL is most robust to both outcome and propensity score model misspecification in the presence of non-overlap.

A key contribution of our work is the introduction of a tuning parameter used to inflate the variance of causal effect estimates in regions of poor data support, so that this variance adequately reflects the high estimation uncertainty in such regions. In simulations the resulting credible intervals are much more reliable than those of competing methods in the presence of non-overlap. However, in "simple" scenarios where causal effect trends in the RN are easily predictable based on trends in the RO, this tuning parameter can produce conservative uncertainties.

Although to our knowledge, the use of Gaussian Process Regression (GPR) for causal inference has not previously been discussed in the literature, there are clear connections between the features of BART + SPL and GPR; thus, a formal comparison is warranted. GPR is a traditionally Bayesian nonparametric regression technique that naturally identifies regions of poor data support and inflates uncertainties in those regions. While these features might make it an attractive and parsimonious approach to causal inference with non-overlap, BART + SPL provides the following advantages over GPR that may render it more appealing, particularly to applied scientists: (1) less sensitivity to tuning choices; (2) more intuitive tuning parameters; and (3) greater computational scalability. BART is highly regarded for its consistent performance under default tuning specifications, and, in Section 3.3, BART + SPL also demonstrated little sensitivity to tuning choices. The results of GPR are known to be sensitive to the choice of kernel and tuning parameters. Similarly, GPR's tuning parameters are often difficult to understand as they are embedded in kernel functions within covariance matrices; therefore, tuning typically requires guess-and-check work. BART + SPL involves tuning parameters with straightforward interpretations relating to the definition of adequate data support. Finally, GPR requires manipulation of a $N \times N$ covariance matrix in

each MCMC iteration, making it time consuming or infeasible with large datasets, while BART + SPL is much more scalable.

Our work introduces an exciting direction for methodological developments in the context of causal inference with propensity score non-overlap. For instance other machine learning methods (Cristianini and Shawe-Taylor (2000), Schmidhuber (2015)) may also have properties that make them well suited to handle non-overlap. Moreover, the limitations of BART in high dimensions provide an opportunity for improvement on our method. In the spirit of Bayesian Adjustment for Confounding (Wang, Parmigiani and Dominici (2012)), propensity score model and outcome model variable selection could be accomplished simultaneously which may reduce the negative effect of irrelevant covariates on BART + SPL. Finally, theoretical results remain to be developed. These could be avenues for future work.

SUPPLEMENTARY MATERIAL

Sampling details, additional simulations, and supplementary tables and figures (DOI: 10.1214/18-AOAS1231SUPP; .pdf). Section 1 of the Supplementary Materials contains a step-by-step description of the BART + SPL MCMC sampling scheme. Section 2 describes the data and results from simulations to test the performance of BART + SPL for binary outcomes. Section 3 provides the supplementary tables and figures referenced in the text.

REFERENCES

AUSTIN, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46** 399–424.

BONATO, V., BALADANDAYUTHAPANI, V., BROOM, B. M., SULMAN, E. P., ALDAPE, K. D. and DO, K.-A. (2011). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics* **27** 359–367.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172

COLE, S. R. and HERNÁN, M. A. (2008). Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.* **168** 656–664.

CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge Univ. Press, Cambridge.

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

D'AMOUR, A., DENG, P., FELLER, A., LEI, L. and SEKHON, J. (2017). Overlap in observational studies with high-dimensional covariates. Available at arXiv:1711.02582.

DI, Q., DAI, L., WANG, Y., ZANOBETTI, A., CHOIRAT, C., SCHWARTZ, J. D. and DOMINICI, F. (2017a). Association of short-term exposure to air pollution with mortality in older adults. *JAMA* **318** 2446–2456.

DI, Q., WANG, Y., ZANOBETTI, A., WANG, Y., KOUTRAKIS, P., CHOIRAT, C., DOMINICI, F. and SCHWARTZ, J. D. (2017b). Air pollution and mortality in the medicare population. *N. Engl. J. Med.* **376** 2513–2522.

FINKEL, M. L. (2016). Shale gas development and cancer incidence in southwest Pennsylvania. *Public Health* **141** 198–206.

GOLDING, B. and WATSON, W. (1999). Possible mechanisms of carcinogenesis after exposure to benzene. *IARC Sci. Publ.* **150** 75–88.

GUTMAN, R. and RUBIN, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Stat. Med.* **34** 3381–3398. MR3412639

HAHN, P. R., MURRAY, J. and CARVALHO, C. M. (2018). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Preprint. Available at arXiv:1706.09523v2.

HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546

HILL, J. and SU, Y.-S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann. Appl. Stat.* **7** 1386–1420. MR3127952

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15** 199–236.

KASSOTIS, C. D., TILLITT, D. E., LIN, C.-H., MCELROY, J. A. and NAGEL, S. C. (2016). Endocrine-disrupting chemicals and oil and natural gas operations: Potential environmental contamination and recommendations to assess complex environmental mixtures. *Environ. Health Perspect.* **124** 256.

KINDO, B. P., WANG, H. and PEÑA, E. A. (2016). Multinomial probit Bayesian additive regression trees. *Stat* **5** 119–131. MR3484083

KING, G. and ZENG, L. (2005). The dangers of extreme counterfactuals. *Polit. Anal.* **14** 131–159.

KLOCZKO, N. (2015). Summary on compressor stations and health impacts. https://www.environmentalhealthproject.org/sites/default/files/assets/downloads/a-brief-review-of-compressor-stations-11.2015.pdf.

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* **113** 390–400. MR3803473

LI, F. and THOMAS, L. E. (2018). Addressing extreme propensity scores via the overlap weights. *Am. J. Epidemiol.* **kwy201**.

LINERO, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636. MR3832214

LIU, Y., TRASKIN, M., LORCH, S. A., GEORGE, E. I. and SMALL, D. (2015). Ensemble of trees approaches to risk adjustment for evaluating a hospital's performance. *Health Care Manage. Sci.* **18** 58–66.

MALTONI, C., CILIBERTI, A., COTTI, G., CONTI, B. and BELPOGGI, F. (1989). Benzene, an experimental multipotential carcinogen: Results of the long-term bioassays performed at the Bologna Institute of Oncology. *Environ. Health Perspect.* **82** 109.

MCKENZIE, L. M., ALLSHOUSE, W. B., BYERS, T. E., BEDRICK, E. J., SERDAR, B. and ADGATE, J. L. (2017). Childhood hematologic cancer and residential proximity to oil and gas development. *PLoS ONE* **12** e0170423.

MESSERSMITH, D., BROCKETT, D. and LOVELAND, C. (2015). Understanding natural gas compressor stations. *Penn State Extension*.

MOKDAD, A. H., DWYER-LINDGREN, L., FITZMAURICE, C., STUBBS, R. W., BERTOZZI-VILLA, A., MOROZOFF, C., CHARARA, R., ALLEN, C., NAGHAVI, M. et al. (2017). Trends and patterns of disparities in cancer mortality among US counties, 1980–2014. *JAMA* **317** 388–406.

NETHERY, R. C, MEALLI, F. and DOMINICI, F. (2019). Supplement to "Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality." DOI:10.1214/18-AOAS1231SUPP.

OAK RIDGE NATIONAL LABORATORY (2017). Natural gas compressor stations. https://hifld-dhs-gii.opendata.arcgis.com/datasets/fd7d62905d194eba87d2ee18d1a244b3_0. Accessed: 2018-03-29.

PELLEGRITI, G., FRASCA, F., REGALBUTO, C., SQUATRITO, S. and VIGNERI, R. (2013). Worldwide increasing incidence of thyroid cancer: Update on epidemiology and risk factors. *J. Cancer Epidemiol*. **2013**.

PENNSYLVANIA DEPARTMENT OF ENVIRONMENTAL PROTECTION (2010). Southwestern Pennsylvania Marcellus Shale short-term ambient air sampling report. http://www.dep.state.pa.us/dep/deputate/airwaste/aq/aqm/docs/Marcellus_SW_11-01-10.pdf. Accessed: 2018-03-29.

PETERSEN, M. L., PORTER, K. E., GRUBER, S., WANG, Y. and VAN DER LAAN, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21** 31–54. MR2867537

R CORE TEAM (2016). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

ROBERT WOOD JOHNSON FOUNDATION (2014). 2014 county health rankings and roadmaps. Prepared by Social Explorer. Accessed: 2018-03-27.

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

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

SCHMIDHUBER, J. (2015). Deep learning in neural networks: An overview. *Neural Netw.* **61** 85–117.

SOUTHWEST PENNSYLVANIA ENVIRONMENTAL HEALTH PROJECT (2015). Summary on compressor stations and health impacts. http://www.environmentalhealthproject.org/files/Summary%20Compressor-station-emissions-and-health-impacts-02.24.2015.pdf. Accessed 2018-03-27.

SPARAPANI, R. A., LOGAN, B. R., MCCULLOCH, R. E. and LAUD, P. W. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Stat. Med.* **35** 2741–2753. MR3513715

US CENSUS BUREAU (2014). American community survey 2014 (5 year estimates). Prepared by Social Explorer. Accessed: 2018-03-27.

US EPA (2018). Endocrine disruptor screening program (EDSP) estrogen receptor bioactivity. https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-program-edsp-estrogen-receptor-bioactivity#main-content. Accessed: 2018-03-27.

WANG, C., PARMIGIANI, G. and DOMINICI, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68** 661–686. MR3055168

WANG, C., DOMINICI, F., PARMIGIANI, G. and ZIGLER, C. M. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics* **71** 654–665. MR3402601

WESTREICH, D. and COLE, S. R. (2010). Invited commentary: Positivity in practice. *Am. J. Epidemiol.* **171** 674–677.

WESTREICH, D., LESSLER, J. and FUNK, M. J. (2010). Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J. Clin. Epidemiol.* **63** 826–833.

WOLF EAGLE ENVIRONMENTAL (2009). Town of Dish, Texas ambient air monitoring analysis final report. https://townofdish.com/objects/DISH_-_final_report_revised.pdf. Accessed: 2018-05-08.

WORLD TRADE CENTER HEALTH PROGRAM (2015). Minimum latency & types or categories of cancer. https://www.cdc.gov/wtc/pdfs/WTCHP-Minimum-Cancer-Latency-PP-01062015.pdf. Accessed: 2018-03-29.

YANG, S. and DING, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika* **105** 487–493. MR3804418

R. C. NETHERY
F. DOMINICI
DEPARTMENT OF BIOSTATISTICS
HARVARD T.H. CHAN SCHOOL OF
PUBLIC HEALTH
BOSTON, MASSACHUSETTS 02115
USA
E-MAIL: rnethery@hsph.harvard.edu
fdominic@hsph.harvard.edu

F. MEALLI
DEPARTMENT OF STATISTICS,
INFORMATICS, APPLICATIONS
UNIVERSITY OF FLORENCE
FLORENCE 50134
ITALY
E-MAIL: mealli@disia.unifi.it