University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2013 *Paper* 309

An Application Of Machine Learning Methods To The Derivation Of Exposure-Response Curves For Respiratory Outcomes

Ekaterina Eliseeva* Alan E. Hubbard†
Ira B. Tager‡

*University of California - Berkeley, Division of Biostatistics, katia.eliseeva@gmail.com
†University of California - Berkeley, Division of Biostatistics, hubbard@berkeley.edu
‡University of California - Berkeley, Division of Epidemiology, ibt@berkeley.edu

An Application Of Machine Learning Methods To The Derivation Of Exposure-Response Curves For Respiratory Outcomes

Ekaterina Eliseeva, Alan E. Hubbard, and Ira B. Tager

Abstract

Analyses of epidemiological studies of the association between short-term changes in air pollution and health outcomes have not sufficiently discussed the degree to which the statistical models chosen for these analyses reflect what is actually known about the true data-generating distribution. We present a method to estimate population-level ambient air pollution (NO2) exposure-health (wheeze in children with asthma) response functions that is not dependent on assumptions about the data-generating function that underlies the observed data and which focuses on a specific scientific parameter of interest (the marginal adjusted association of exposure on probability of wheeze, over a grid of possible exposure values). We show that this approach provides a more nuanced summary of the data than more typical statistical methods used in air pollution epidemiology and epidemiological studies in general.

Introduction

Quantitative estimates of the effects of ambient air pollutants on human health are derived almost exclusively from epidemiological studies. A voluminous literature has been generated to provide estimates of the impacts of short and long-term exposures. Time series studies have been the mainstay for assessment of short-term associations ([1],[2],[3],[4],[5],[6]), and cohort and cross-sectional studies provide estimates of associations related to exposures over the course of years ([7],[8],[9]).

In addition to emission source strengths, ambient air pollutant concentrations are highly dependent on seasonal characteristics of meteorological factors (e.g. temperature, humidity, depth of mixing layer, etc. ([10])). Similar factors affect a broad range of health and physiological outcomes (e.g. heart attacks, respiratory illnesses, hospitalizations for a variety of diseases) ([11],[12],[13],[14]). Consequently, considerable attention has been paid to the control of these temporal, potentially confounding factors, in particular, meteorological variables ([15],[16],[17]).

In contrast to the attention paid to the functional forms of the confounders, much less work has been carried out in defining the exposure-response relation between pollutants and outcomes. Most studies have used either parametric or semi-parametric models, with the latter often still constrained by strong assumptions, like additivity of the risk factors. Some have parameterized the exposure coefficient as a single linear term, as unconstrained or constrained lag functions in time series, or as simple linear terms in chronic exposure studies ([18],[19],[20],[21],[1],[17]). While efforts have been made to characterize the shapes of the exposure-response associations, most study's analyses have been carried out within the framework of a single class of models, thus making potentially strong assumptions of the joint functional form of confounders and the exposure variable ([22],[23]). Characterizations of the functional form of the exposure-response relation have used linear models, parametric threshold models ([24],[25]), polynomials ([26]), natural cubic splines ([22],[5]), penalized

splines ([8]), or have been implied by comparison of a linear model with and without a truncated distribution of exposure at a specific "threshold" value ([20]).

A related issue concerns the type of parameter typically estimated by existing studies. Specifically, given the estimates are simply direct byproducts of the models used, estimates from most time series and cohort analyses are so-called conditional estimates, i.e., the interpretation of the coefficient(s) in front of the air pollution term (or the "smoothed" exposure-response") is conditional on fixing the other covariates in the model. Interpretation of this "conditional" association becomes problematic in the almost-certain case that those highly constrained models are misspecified. Generally, this fact is ignored and the coefficient(s) are treated as if they were marginal (in terms of being averaged association over the entire population distribution of covariates/confounders, such as the average treatment effect), the one exception being the type of Bayesian Model Averaging (BMA) used by Schwartz, et al. ([8]). This has serious consequences when relative risk/hazard (or relative odds) are treated as if they were true population expected values ([27]) and are then incorporated into risk assessments to estimate population burden ([28],[29]) or to estimate change in population burden with changes in air pollutant concentrations over time ([31],[32]).

This paper addressed these issues simultaneously, in that 1) we discuss estimation of the model of the outcome given an air pollutant in a very "big" regression model, meaning one that assumes very little about the exposure-response, or statistical interactions, for instance, and 2) defines an estimator that does not depend on the form of the regression model, and leverages this fit to estimate a parameter with a direct public health interpretation: a marginally adjusted risk of health outcomes. Specifically, we describe the use of Super Learner (SL) - a cross-validation based, ensemble-learner approach to prediction, that casts a very large net for fitting models, without over-fitting them. We apply this approach to obtain estimates of the daily probability of wheezing had the entire population of children in the Fresno Asthmatic Children's Environment Study (FACES) been exposed to specific

levels of NO_2 (two days prior to measuring wheezing) conditional on a high-dimensional vector of assumed confounders. We discuss a straightforward use of Super Learner (SL) for estimating the standardized, marginal risk of wheezing over a grid of potential values for NO_2 (a method equivalent to the G-computation approach of Robins ([30])). Under (strong) causal assumptions of conditional exchangeability, consistency, and positivity, these treatment-level specific standardized risks may be interpreted causally- specifically as the probability of wheeze on a given day had NO_2 two days before been set to that particular exposure level. This example shows that the method we propose does as well as the typically used parametric approach.

Materials and Methods

The approach we present derives from two majors issues, which are distinct but important components of research in causal inference. First, we want to be able to model the data-generating distribution, i.e., the system that produced the FACES data, constraining our model only by information we actually have. Specifically, we have an idea of the process that produces the health outcomes (wheeze) via exposure (NO_2) and covariates (Table 1), but one rarely, if ever, has firm theoretical justification for the multivariate (confounders+variable of interest) response of the outcome of interest. Optimally, one would define the estimating model to be one that follows only the known constraints of the mechanism by which the data is generated, which are typically minimal, or a nearly nonparametric models. However, given the so-called 'curse of dimensionality', nonparametric estimation is impossible (in reasonable sample sizes, there is rarely enough observations in every unique group defined by the predictor variables to get a reasonable estimate of the mean within such groups), so some compromise must be made, and that compromise is estimation in a large semi-parametric model. Thus, instead of a compromise that is based on an arbitrary model, we discuss below a data-adaptive method (SL) that finds the "best" model in a huge class of potential models.

Because the model is unknown, putting constraints on it in order to get it to return an interpretable estimated parameter must lead to bias of unknown magnitude. However, SL returns a regression with no such convenient form (nor should it), so one needs estimates of a parameter of interest that is simple and meaningful to interpret, but is not tied to a specific model. Thus, we estimate parameters (standardized risks) that do not rely on the statistical specification of a particular parametric or semi-parametric model, but instead rely on specific questions (e.g., how many children would wheeze on average if they all experienced a particular value of NO_2).

Data

The data come from the Fresno Asthmatic Children's Environment Study (FACES) ([43]). The goal of this study was to examine the effects of air pollution on children with asthma. Briefly, a convenience sample of children with asthma was recruited between 2000 and 2005, and data collection ran from 2000 to 2008 ([44]). Eligibility criteria were ages 6-11 at the start of the study, a physician's diagnosis of asthma, having active asthma (indicated by use of asthma medication, asthma symptoms, or asthma-related healthcare utilization), living in their primary residence for at least 3 months prior to enrollment in the study and no plans to move for the next 2 years, living within 20 km of the US Environmental Protection Agency (EPA) air quality monitoring site located in Fresno, CA, and no physical or mental conditions that could impair completion of the study protocol ([45]).

We define the unit of observation as the child, i, who has measurements of variables on each day, j, including their outcome Y_{ij}, as well as time-dependent exposure, A_{ij}, and a vector of adjustment variables, W_{ij}. For each child there is also a missing indicator for the outcome, Δ_{ij}. Thus we observe i.i.d. occurrences of $O_i = (\Delta_i, \Delta_i Y_i, A_i, W_i)$, where, for example, $A_i = (A_{i1}, A_{i2}, ..., A_{im})$. We focused on estimating the relationship between the probability of wheezing for each child (binary outcome), and exposure to NO_2 (measured

in ppb), (a continuous exposure), while controlling for a number of covariates. During each panel day, the presence or absence of wheeze was recorded between 6 and 9 AM and responses to questions were programmed into the portable spirometer ("Did you wheeze after bedtime?"). NO_2 data were collected from a central site monitor in Fresno, CA, and exposures were assigned for each child on the day the measurement was taken. Thus, all children received the same air pollutant exposure assignment, if they were in the same two-week panel. To account for a temporal relationship between NO_2 and wheeze, lags and moving averages were created for days 1 through 14 prior to the outcome assessment. To illustrate our method, we focus only on lag one of NO_2 in our analysis (NO_2 measured in the 48 hour period before the test). Covariates included those measured at baseline, at each 6-month or annual field office visit, and environmental covariates such as temperature (Table 1). Thee confounders were chosen and justified in a previous paper ([44]. There were a total of 15252 observations with non-missing outcomes among 280 children.

Statistical Methodology

Parameter of Interest

Our parameter of interest, is the marginally adjusted association of exposure (NO_2) on probability of wheeze, adjusting for a large set of confounding variables, or explicitly, over a grid of possible exposure values, a:

$$\theta(a) \equiv E_W \left[E(Y|(A = a, W) \right]$$

If the assumptions of positivity and experimental treatment assignment hold, this parameter is equal to $E_W(Q_0(a, W))$, where $Q_0(a, W)$ is the true conditional expectation function. In other words, we want the estimated mean, over the observed distribution of the W in our population (covariates measured over all children, all days in our study). If we knew the true conditional mean ($Q_0(a, W)$), and could replace the observed values of exposure over

time and children (the A) to a single value, a, we could examine this value (say via a plot) over a grid, from relatively low to high values of fixed exposure, a, or $\hat{\theta}(a)$ vs a.

Without further assumptions, this is a reasonable way to get a standardized risk, but it has an even more appealing interpretation under stronger assumptions: $\theta(a) = E_W(Y(a))$. This is is an estimate of one-day lagged marginal probability of wheezing for every day of the study for all subject had $NO_2 = a$, $P(Y(a) = 1)$, where $Y(a)$ represents the outcome for a particular subject on a particular day if he had, contrary to fact, been exposed to level a of NO_2 two days previously.

Estimation Procedure

First, we need an estimate of the regression $Q_0(A, W)$, and as discussed above, we have the challenge of many covariates (high dimensional data), but in a huge statistical model. We use the Super Learner (SL), an ensemble, machine-learning algorithm ([41]), which allows for the simultaneous evaluation (by cross-validation) of a library of plausible model algorithms (including potentially the investigators' a priori chosen model) to determine which of the models are most appropriate for the data, based on minimizing a least squares loss function, and then averages over these chosen models to produce a composite model. This process makes few assumptions about the relation between wheeze and the joint distribution of NO_2 and confounders. The SL performs asymptotically as well (in terms of expected risk difference) as the oracle estimator (the estimator that comes closest to the truth if it were known), up to a second order term. While the number of learning algorithms considered by SL is polynomial in sample size, the SL is relatively optimal because it converges to the oracle selector if none of its candidate learners (and oracle estimator) converges at a parametric rate. If one of the parametric candidate learners is actually the true model, however, and thus converges at a parametric rate, the SL will converge at almost the parametric rate of $\log n/n$ ([41]). Thus the SL theory encourages the use of a very large number of possible learning algorithms with very little penalty if the true model is one of the simple, parametric

models contained in the set of learners.

We use a simple substitution estimator, sometimes called the G-computation algorithm ([30]) in the context of point treatments (that is, we are not treating every day as a serial cross-sectional study) and our estimating model, $Q(a, W)$ equals the true model, $Q_0(a, W)$, assuming our assumption of positivity holds. Our estimator is

$$\hat{\theta}(a) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{Q}(a, W_{ij}),$$

and can be obtained by first setting A to values a in the estimated model $Q(a, W_{ij})$ for each child, while keeping their covariates at their observed values. Then these estimates would be averaged over all children for each level of a. Note that we are assuming that we can extrapolate our SL curve to estimate $\hat{\theta}(a)$ even for those children who may never have experienced the set exposure level. However we did test for positivity (the amount of extrapolation) and found no significant violations ([55]). Also note that we predicted outcomes for all observations that a child was supposed to have had, even if they were originally missing a panel day. Thus we were predicting for 32189 observations rather than 15252.

For comparison, we also plugged in averages of covariates into our SL fit to predict outcomes. Thus we were calculating $E(Y|A = a, W = \bar{w})$ over a grid of a. This would mimic the typical way of generating an "adjusted" curve.

We also fit unadjusted models, i.e., models with only NO_2 as a predictor of the outcome, using the SL and a main-terms only generalized additive model (GAM) with a spline option and 4 degrees of freedom, the last because it has been frequently used in time series analyses of associations between daily changes in air pollution and a variety of health effects.

The R ([49]) package SuperLearner ([41]) were used to run the SL machine learning algorithm. Since the dataset consisted of repeated measures on each child, an "ID" variable was used to

make sure that the V-fold cross-validation splits kept observations from the same individuals in the same split. The binomial family was also specified for both the SuperLearner and GAM models.

To handle missing covariate, W_{ij}, we redefined the data to include $\delta_i j$ and $\delta_{ij} W_{ij}$, where $\delta_{ij} = 1$ if the covariate was observed and 0 otherwise. The SL library contained a generalized additive model (GAM) ([50]), a GAM with 3 degrees of freedom, generalized linear models (GLM) ([51]), neural networks (NNET) ([52]), NNET with tuning parameter set to 3, GLMNET, and the GAM with spline model with NO_2 as the only predictor and four degrees of freedom as candidate learners (Table 4). These learning algorithms were chosen because they vary in the way they perform model selection. Specifically, when these are combined they have the potential for and the flexibility with which they can fit a regression, but as well contain learners that are highly parametric (single regression) if such models are better fits. Additional tuning parameters for the SL were the number of folds in V-fold cross-validation, set to 5, and a loss function, which was the non-negative least squares loss function.

Since the model for the average outcome, given observed exposures and confounders, was estimated data-adaptively based on a (nearly) unspecified (semi-parametric) model, asymptotically this method should converge, as sample size grows, to a consistent estimate of $\theta(a)$ and, thus, the true exposure-response curve.

Note that given our model is (nearly) non-parametric, our estimator is not an asymptotically linear estimator ([46]), and we cannot rely on the asymptotic normality of the pointwise estimates, $\hat{\theta}(a)$, over a grid on a. However, we can calculate uncertainty intervals akin to pointwise confidence intervals (CIs) at each exposure level using the nonparametric bootstrap procedure ([53]). Because we had repeated measures data, and the children were assigned randomly to the seasonal, 14-day panels, we performed a "clustered" bootstrap with the children as the cluster ([54]). We refit the model for $Q(A, W)$, using the SL algorithm on each

sampled dataset and used each fit of the model to make predictions on the bootstrap samples. The substitution estimator of the G-computation formula was then used to recalculate the exposure-response curve for the range of NO_2 observed in the data. The 2.5th and 97.5th percentile of the bootstrap distribution were selected at each level of NO_2 to obtain 95% confidence intervals. Note we can also predict Y's for subjects over each of the panels in which they were enrolled in the study, as we still observe the $A_{ij} = a, W_{ij}$ and thus we average over the entire complete data. This implicitly assumes that the data are missing at random (MAR) ([59]).

We also calculated the 95% confidence interval for the difference between the estimator at NO_2 levels of 35 and and 5 (an estimate of $E_w\{Q(35, W) - Q(5, W)\}$) as well as the relative risk (an estimate of $E_w\{Q(35, W)/Q(5, W)\}$) to see whether there was a significant difference in the prediction between these two levels.

To see how well our models were predicting we split our data into 10 groups and ran SL and GAM on each group separately, while predicting on the remaining groups. We then plotted ROC curves based on these cross-validated predictions, to get an unbiased estimate of performance.

Results

Descriptive statistics for Y, A, and 36 covariates are in Table 2. Figures 1a-1b present the estimated unadjusted curves for predicted probability of wheeze for GAM with spline and SL.

The red line in Figure 2 displays the estimate of the marginal exposure-response curve obtained from GAM with a smoothing spline. Its shape is similar to the equivalent unadjusted curve based on GAM (essentially a bivariate smoothing spline).

Of all the candidate learners considered in the SL algorithm, only two were given a non-

zero weight in the final model (weights must sum to one). The weights for these learners were: 0.74 for the GAM with spline model and 0.26 for the GLM model. It is interesting to note that most of the weight in the SL was given to the preferred model used in air pollution research. However, note that our model indicates that the relative difference is approximately linear, not the relative risk. The estimated marginal exposure-response curve obtained from the SL is in (Figure 2, black line). This curve is relatively flat over the range of the observed NO_2. The scale of probabilities on the vertical axis indicates that this curve shows is steeper than from the adjusted GAM. We note that the confidence intervals from the GAM model are narrower than the ones from SL, which is what one would expect if the best choice (the oracle selector ([41])) was in fact GAM However, as one can see, there is little loss in precision in the SL, even with a much bigger model, as the oracle inequality predicts.

The estimate of $E_w\{Q(35, W) - Q(5, W)\}$ for the SL was 0.067 and the 95% confidence limits were (0.032, 0.094). The estimate for GAM was 0.043 and the 95% confidence interval was (0.011, 0.082). Thus, the SL estimate is 50% larger than the one from GAM, though both indicate a significant change for higher NO_2 vs. a lower exposure.

The estimate of $E_w\{Q(35, W)/Q(5, W)\}$ for the SL was 1.53 and the 95% confidence limits were (1.27, 1.98). The estimate for GAM was 1.30 and the 95% confidence interval was (1.07, 1.65).

Finally, we show the ROC curves (Figure 3) of the cross-validated predictors based on GAM and SL and see that neither algorithm predicts wheezing very accurately.

Discussion

Epidemiological studies of the association between short-term changes in air pollution and health outcomes have not generally discussed the degree to which the statistical models

chosen for analyses of exposure-response functions reflect what is actually known about the true underlying data-generating distribution. Many analyses have applied statistical models without first carefully focusing on the scientific parameter of interest - increase in health outcome per unit of increased exposure. Emphasis has been on relative changes in outcome or survival rather than on direct estimates of this parameter. The emphasis on relative changes has a long history derived from Cornfield ([56]); the consequence is that absolute estimates of disease burden have not been used as much as they might have been. The relative association (effect) parameter is often simply a byproduct of a statistical procedure, such as Poisson regression, not chosen a priori to directly address the scientific question of interest, which is to estimate an exposure-response function from which one can derive an estimate of health hazard for a given level of exposure. Such models provide only relative estimates of hazard without any regard to the actual underlying ("baseline") risk, which is of importance in and of itself. Finally, estimation has often been done under the implicit assumption that much more is known about the data-generating distribution than can actually be supported by previous work or theory. The approach we employed defines a parameter of interest to environmental epidemiologists– an adjusted, marginal exposure-response curve; it does not presume knowledge of the true data-generating distribution (i.e., the model is semi-parametric) and we can derive standardized risk and/or relative risk estimates from it.

Results based on a composite model derived from a library of different model-fitting algorithms showed an exposure-response function for ambient NO_2 concentrations in the 48 hours prior to assessment of wheeze status in a group of children with asthma that is most consistent with no threshold (Figure 2).

The results (estimates and inference) from our SL approach were not very different from those of the approach using a smaller statistical model (GAM). This fact might seem to be an advertisement for simpler methods. However, because it is never known a priori what the functional regression form is, we are only able to know this in hindsight. In this

case, estimation with the SL and the resulting inference is consistent with the lack of a priori knowledge about the statistical model. The fact that the oracle inequality states the potential gain in accuracy by using SL with a large library, and the relatively low cost of fitting a big model even when a simpler model is "true", one essentially gets to have their cake and eat it too. This is underscored by the similar range of the 95% CI comparing estimates of the population-level change in probability of wheezing comparing high to low levels of exposure.

We have not provided a complete analysis of the response functions for various lags or types of moving averages. Instead, we have focused on the shape of the curve and its interpretation relative to other methods. In addition, we used data from a fixed central site monitor to define exposure, which we acknowledge would not be the optimal method if the goal were to focus specifically on the relation of the exposure and wheeze in our population, which would require individual-level measurements of NO_2 for each child. Individual-level exposure estimates are available to us and will be used in analyses focused on specific health outcomes.

In summary, we present a method to estimate population-level ambient air pollution exposure-response functions that are not dependent on assumptions about the underlying data-generating distribution, which focuses on a specific scientific parameter of interest (risk difference), and which, from a theoretical standpoint, asymptotically provides an estimator that is optimal with respect to variance and bias given a large library of learners.

Bibliography

[1] *Revised Analysis of Time-Series Studies of Air Pollution and Health-Special Report*, (Boston MA), Health Effects Institute, 2003.

[2] J. Schwartz and A. Marcus, "Mortality and air pollution in london: a time series analysis," *Am J Epidemiol*, vol. 131, pp. 185–194, Jan 1990.

[3] J. Schwartz and A. Zanobetti, "Using meta-smoothing to estimate dose-response relationship trends across multiple studies, with application to air pollution and daily death," *Epidemiology*, vol. 11, pp. 666–672, 2000.

[4] R. J. Delfino, H. Gong, W. S. Linn, Y. Hu, and E. Pellizzari, "Respiratory symptoms and peak expiratory flow in children with asthma in relation to volatile organic compounds in exhaled breath and ambient air," *J Expo Anal Enivron Epidemiol*, vol. 13, pp. 348–363, Sep 2003.

[5] J. Zhou, K. Ito, R. Lall, M. Lippmann, and G. Thurston, "Time-series analysis of mortality effects of fine particulate matter components in detroit and seattle," *Environ Health Perspect*, vol. 119, pp. 461–466, 2011.

[6] B. Ostro, M. Lipsett, J. Mann, A. Krupnick, and W. Harrington, "Air pollution and respiratory symptoms among adults in southern california," *Am J Epidemiol*, vol. 137, pp. 691–700, 1993.

[7] D. K. D., M. Jerrett, R. Burnett, R. Ma, E. Hughes, Y. Shi, M. Turner, C. Pope, G. Thurston, E. Calle, and M. Thun, eds., *Extended Follow-up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality*, no. 140, (Boston), Health Effects Institute, 2009.

[8] J. Schwartz, B. Coull, F. Laden, and L. Ryan, "The effect of dose and timing of dose on the association between airborne particles and survival," *Environ Health Perspect*, vol. 116, pp. 64–69, Jan 2008.

[9] N. Kunzli, M. Jerrett, W. J. Mack, B. Beckerman, L. LaBree, F. Gilliland, D. Thomas, J. Peters, and H. Hodis, "Ambient air pollution and atherosclerosis in los angeles," *Environ Health Perspect*, vol. 113, pp. 201–206, Feb 2005.

[10] G. McGregor, "Basic meteorology," in *Air Pollution and Health* (S. T. Holgate, J. M. Samet, H. S. Koren, and R. L. Maynard, eds.), Academic Press, 1999.

[11] M. O'Neill, A. Zanobetti, and J. Schwartz, "Modifiers of the temperature and mortality association in seven u.s. cities," *Am J Epidemiol*, vol. 157, pp. 1074–1082, 2003.

[12] P. Goodman, D. W. Dockery, and L. Clancy, "Cause-specific mortality and the extended effects of particulate pollution and temperature exposure," *Environ Health Perspect*, vol. 112, pp. 179–185, 2003.

[13] L. Moseholm, E. Taudorf, and A. Frosig, "Pulmonary fucntion changes in asthmatics associated with low-level so2 and no2 air pollution, weather, and medicine intake," *Allergy*, vol. 48, pp. 334–344, 1993.

[14] R. Basu and B. D. Ostro, "A multicounty analysis identifying the populations vulnerable to mortality associated with high temperature in california," *Am J Epidemiol*, vol. 168, pp. 632–637, 2008.

[15] L. J. Welty and S. L. Zeger, "Are the acute effects of particulate matter on mortality in the national morbidity, and mortality, and air pollution study the result of inadequate control for weather and season? a sensitivity analsyis using flexible distributed lag models," *Am J Epidemiol*, vol. 162, pp. 80–88, 2005.

[16] M. S. Goldberg and R. T. Burnett, eds., *Revised Analysis of the Montreal Time-Series Study*, (Boston), Health Effects Institute, 2003.

[17] K. Ito, R. Mathes, Z. Ross, A. Nadas, G. Thurston, and T. Matte, "Fine particulate matter constituents associated with cardiovascular hospitalizations and mortality in new york city," *Environ Health Perspect*, vol. 119, pp. 467–473, 2011.

[18] R. J. Delfino, N. Staimer, T. Tjoa, D. Gillen, M. T. Kleinman, C. Sioutas, and D. Cooper, "Personal and ambient air pollution exposures and lung function decrements in children with asthma," *Environ Health Perspect*, vol. 116, pp. 550–558, 2008.

[19] S. L. Zeger, F. Dominici, A. McDermott, and J. Samet, "Mortality in the medicare population and chronic exposure to fine particulate air pollution in urban centers (2000-2005)," *Environ Health Perspect*, vol. 116, pp. 1614–1619, Dec 2008.

[20] M. Jerrett, R. T. Burnett, C. A. Pope, K. Ito, G. Thurston, D. Krewski, Y. Shi, E. Calle, and M. Thun, "Long-term ozone exposure and mortality," *New England Journal of Medicine*, vol. 360, pp. 1085–1095, 2009.

[21] R. Lall, K. Ito, and G. Thurston, "Distributed lag analysis of daily hospital admissions and source-apportioned fine particle air pollution," *Environ Health Perspect*, vol. 119, pp. 455–460, 2011.

[22] M. J. Daniels, F. Dominici, J. M. Samet, and S. L. Zeger, "Estimating particulate matter-mortality dose-response curves and threshold levels: An analysis of daily time-series for the 20 largest us cities," *Am J Epidemiol*, vol. 152, pp. 397–406, 2000.

[23] J. Schwartz, F. Laden, and A. Zanobetti, "The concentration-response relation between pm2.5 and daily deaths," *Environ Health Perspect*, vol. 110, pp. 1025–1029, 2002.

[24] F. Dominici, M. Daniels, A. McDermott, S. Zeger, and J. L. Samet, "Shape of the exposure-response relation and mortality displacement in the nmmaps study," in *Special Report: Revised Analyses of Time-Series Studies of Air Pollution and Health*, pp. 91–96, Charlestown, MA: Health Effects Institute, 2003.

[25] *Extended Follow-up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality*, (Boston), Health Effects Institute, 2009.

[26] B. Ostro, "A search for a threshold in the relationship of air pollution to mortality: A reanalysis of data on london winters," *Environ Health Perspect*, vol. 58, pp. 397–399, Dec 1984.

[27] S. Cakmak, R. T. Burnett, and D. Krewski, "Methods for detecting and estimating population threshold concentrations for air pollution-related mortality with exposure measurement error," *Risk Anal.*, vol. 19, pp. 487–496, 1999.

[28] *National Research Council Committee on Improving Risk Analysis Approaches Used by the U.S. EPA. Science and Decisions: Advancing Risk Assessment: Chapter 5: Toward a Unified Approach to Dose-Response Assessments.* Washington, D.C.: National Research Council of the National Academies, 2009.

[29] S. C. Anenberg, L. W. Horowitz, D. Q. Tong, and J. J. West, "An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling," *Environ Health Perspect*, vol. 118, pp. 1189–1195, 2010.

[30] J. Robins, "A new approach to causal inference in mortality studies with a sustained exposure period- application to control of the healthy worker survivor effect," *Mathematical Modeling*, vol. 7, pp. 1393–1512, 1986.

[31] M. L. Bell, D. L. Davis, V. H. Gouveia, N. Borja-Aburto, and L. A. Cifuentes, "The avoiodable health effects of air pollution in three latin american cities: Santiago, sao paulo, and mexico city," *Environ Res.*, vol. 100, pp. 431–440, 2006.

[32] C. Schindler, D. Keidel, M. W. Gerbase, E. Zemp, R. Bettschart, O. Brandll, and M. H. Brutsche, "Improvements in pm10 exposure and reduced rates of respiratory symptoms in a cohort of swiss adults (sapaldia)," *Amer J Resp Crit Care Med.*, vol. 179, pp. 579–587, 2009.

[33] E. Samoli, R. Peng, T. Ramsay, M. Pipika, G. Touloumi, F. Dominici, R. Burnett, A. Cohen, D. Krewski, J. Samet, and K. Katsouyanni, "Acute effects of ambient particulate matter on mortality in europe and north america: Results from aphena study," *Environ Health Perspect*, vol. 116, pp. 1480–1486, 2008.

[34] R. D. Peng, F. Dominici, and T. A. Louis, "Model choice in time series studies of air pollution and mortality," *JR Statist Soc A.*, vol. 169(pt 2), pp. 179–203, 2006.

[35] R. J. Delfino, N. Staimer, T. Tjoa, D. Gillen, A. Polidori, M. Arhami, M. T. Kleinman, N. D. Vaziri, J. Longhurst, and C. Sioutas, "Air pollution exposures and circulating biomarkers of effect in a susceptible population: Clues to potential causal component mixtures and mechanisms," *Environ Health Perspect*, vol. 117, pp. 1232–1238, 2009.

[36] M. L. Bell, A. McDermott, S. L. Zeger, J. M. Samet, and F. Dominici, "Ozone and short-term mortality in 95 us urban communities, 1987-2000," *JAMA*, vol. 292(19), pp. 2372–2378, Nov 2004.

[37] J. I. Halonen, T. Lanki, P. Tiittanen, V. N. Jarkko, M. Loh, and J. Pekkanen, "Ozone and cause-specific cardiorespiratory morbidity and mortality," *J Epidemiol Commun Health*, vol. 64, pp. 814–820, 2010.

[38] M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan, "Positivity," in *Targeted Learning: Causal Inference for Observational and Experimental Data* (M. J. van der Laan and S. Rose, eds.), New York: Springer, 2011.

[39] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

[40] M. J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics, 2011.

[41] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, "Super learner.," *Stat Appl Genet Mol Biol*, vol. 6, p. Article 25, 2007.

[42] J. M. Snowden, S. Rose, and K. M. Mortimer, "Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique," *Am J Epidemiol*, vol. 173, pp. 731–738, 2011.

[43] "http://facesproject.berkeley.edu/."

[44] J. K. Mann, J. R. Balmes, T. A. Bruckner, K. M. Mortimer, H. G. Margolis, B. Pratt, S. K. Hammond, F. Lurmann, and I. B. Tager, "Short-term effects of air pollution on wheeze in asthmatic children in fresno, california," *Environ Health Perspect*, 2010.

[45] H. G. Margolis, J. K. Mann, F. W. Lurmann, K. M. Mortimer, J. R. Balmes, S. K. Hammond, and I. B. Tager, "Altered pulmonary function in children with asthma associated with highway traffic near residence.," *Int J Environ Health Res*, vol. 19, pp. 139–155, Apr 2009.

[46] M. J. van der Laan and J. M. Robins, *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

[47] M. J. van der Laan and D. Rubin, "Targeted maximum likelihood learning," tech. rep., UC Berkeley Division of Biostatistics, 2006.

[48] S. E. Sinisi and M. J. van der Laan, "Deletion/substitution/addition algorithm in learning with applications in genomics.," *Stat Appl Genet Mol Biol*, vol. 3, p. Article18, 2004.

[49] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[50] T. Hastie and R. J. Tibshirani, *Generalized Additive Models*. Chapman and Hall, 1990.

[51] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. Champan and Hall, 1989.

[52] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*. New York: Springer, fourth ed., 2002.

[53] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial Mathematics, 1987.

[54] C. A. Field and A. H. Walsh, "Bootstrapping clustered data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 3, pp. 369–390, 2007.

[55] K. L. Moore, R. S. Neugebauer, M. J. van der Laan, and I. B. Tager, "Causal inference in epidemiological studies with strong confounding," *Stat Med.*, vol. 31(13), pp. 1380-404, 2012.

[56] C. Poole, "On the origin of risk relativism," *Epidemiology*, vol. 21, pp. 3–9, 2010.

[57] J. Schwartz, A. Zanobetti, and T. Bateson, "Morbidity and mortality among elderly residents of cities with daily pm measurements.," in *Special Report: Revised Analyses of Time-Series Studies of Air Pollution and Health.*, pp. 25–58, Charlestown, MA: Health Effects Institute, 2003.

[58] R. Neugebauer and M. J. van der Laan, "Why prefer double robust estimators in causal inference?," *J Stat Plan Inference*, vol. 129, pp. 405–426, 2005.

[59] D.B. Rubin, "Inference and missing data (with discussion).," *Biometrika*, vol. 63, pp. 581–592, 1976.

Table 1: Table of Covariates used in the FACES data analysis

Baseline	Time-varying	Environmental
gender	height	temperature
income category	age	relative humidity
low birth weight	bird/cat/rodent/dog in home	season
father's/mother's asthma status	eczema	.
asthma diagnose before age 2	rhinitis	.
smoking during pregnancy	ownership of home	.
premature birth	household smoking policy	.
race/ethnicity	B[a]/C[b] medication status	.
age at asthma diagnosis	.	.
breastfeeding status	.	.
symptom severity score	.	.
income category	.	.
positive skin tests	.	.

[a] inhaled steroids and intal/cromolyn

[b] not beta-agonists or inhaled steroids

Table 2: Descriptive statistics of the variables in the FACEs dataset

Variable	N	Mean	SD	Min/Max
Wheeze (1=yes, 0=no)	15252	0.15	0.36	0/1[a]
NO_2 Lag 1	15252	19.59	8.10	4.62/52.44
Low Birth Weight	14681	0.08	0.27	0/1[a]
Premature	14801	0.10	0.30	0/1[a]
Breastfed	14875	0.73	0.44	0/1[a]
Male	15252	0.58	0.49	0/1[a]
Income Category	14883	2.72	1.10	1/4
Own Home	13762	0.66	0.48	0/1[a]
Hispanic	15252	0.41	0.49	0/1[a]
Black	15252	0.11	0.31	0/1[a]
White	15252	0.45	0.50	0/1[a]
Age (years)	15252	9.42	1.97	6/14
Height (cm)	15200	139.2	14.1	106.4/182.9
B Meds	15252	0.52	0.50	0/1[a]
C Meds	15252	0.30	0.46	0/1[a]
Bird in Home	15252	0.12	0.33	0/1[a]
Cat in Home	15252	0.25	0.43	0/1[a]
Dog in Home	15252	0.35	0.48	0/1[a]
Rodent in Home	15252	0.13	0.33	0/1[a]
Positive Skin Test	13756	0.65	0.48	0/1[a]
Eczema	15070	0.16	0.37	0/1[a]

Rhinitis	15070	0.13	0.34	0/1[a]
Diagnosis ≤ 2	15006	0.38	0.49	0/1[a]
Age First Diagnosed	15006	3.79	2.70	0/11
Father's Asthma	13271	0.28	0.45	0/1[a]
Mother's Asthma	14861	0.38	0.48	0/1[a]
Symptom Severity Score	15252	1.93	0.70	1/3
Prenatal Smoking	14047	0.08	0.27	0/1[a]
Home No Smoking Policy	14926	0.97	0.18	0/1[a]
Home Smoking	15250	0.18	0.38	0/1[a]
Year	15252	2002.98	1.16	2000/2005
Month	15252	6.47	3.58	1/12
Winter (Nov-Feb)	15252	0.35	0.48	0/1[a]
Spring (Mar-June)	15252	0.33	0.47	0/1[a]
Sumer (July-Oct)	15252	0.32	0.47	0/1[a]
Apparent Temperature	15252	16.28	8.38	0.33/36.66
Avg Temperature (°C)	15252	16.28	7.57	0.33/36.66

[a] 1 indicates yes/present, 0 indicates no/absent

Table 3: SuperLearner Candidate Algorithms

Algorithm	Description
GAM	generalized additive model with 2 df and considers any variable with more than 4 unique values to be continuous and able to be in smoothing splines
GAM	generalized additive model with 3 df and considers any variable with more than 4 unique values to be continuous and able to be in smoothing splines
GLM	generalized linear model with all main terms
NNET	neural net with size (number of units in hidden layer) 1
NNET	neural net with size (number of units in hidden layer) 3
GLMNET	generalized linear model with all main terms with penalized likelihood
GAM w/ spline	generalized additive model with 4 df and NO_2 fixed in the model

Figures

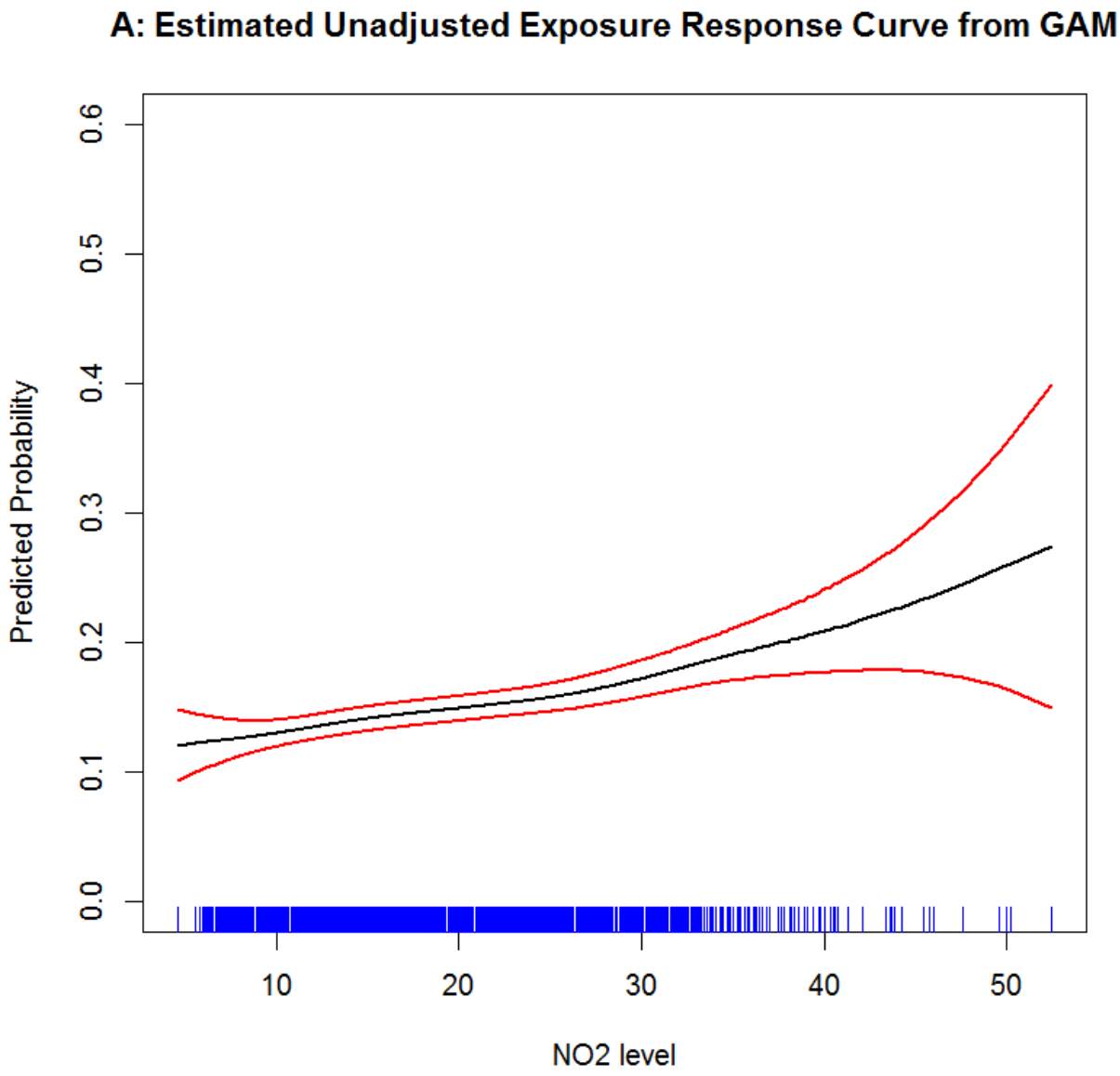

Figure 1: a) Estimated unadjusted marginal exposure-response curve and CIs obtained using GAM with a spline model with 4 df.

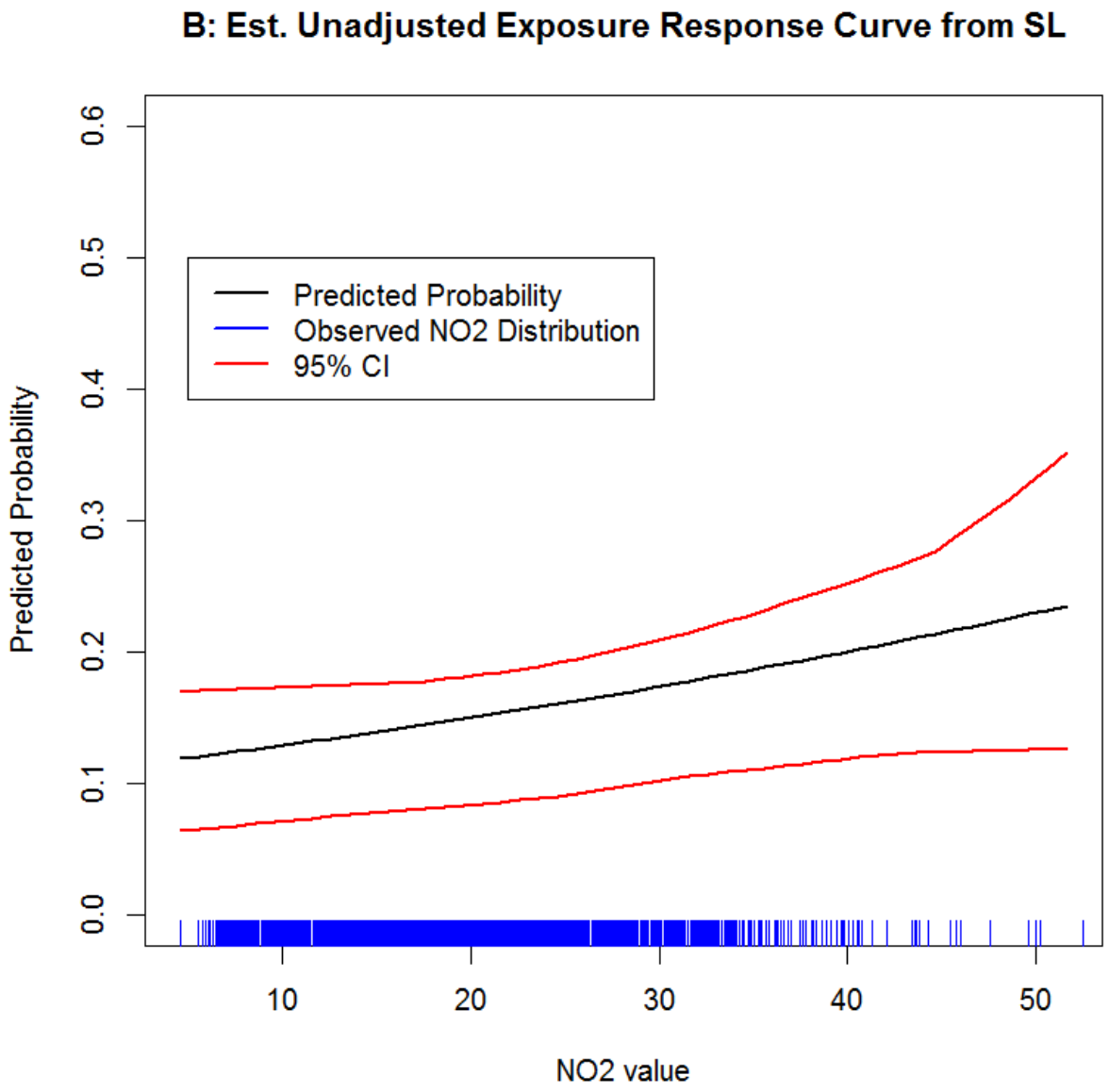

Figure 1: b) Estimated unadjusted marginal exposure-response curve and CIs obtained using SL. The horizontal axis is based on the NO_2 distribution observed in the data. The blue bars on the horizontal axis indicate actual NO_2 levels that were observed. The vertical axis is the predicted probability of wheezing. The red lines are 95% confidence limits.

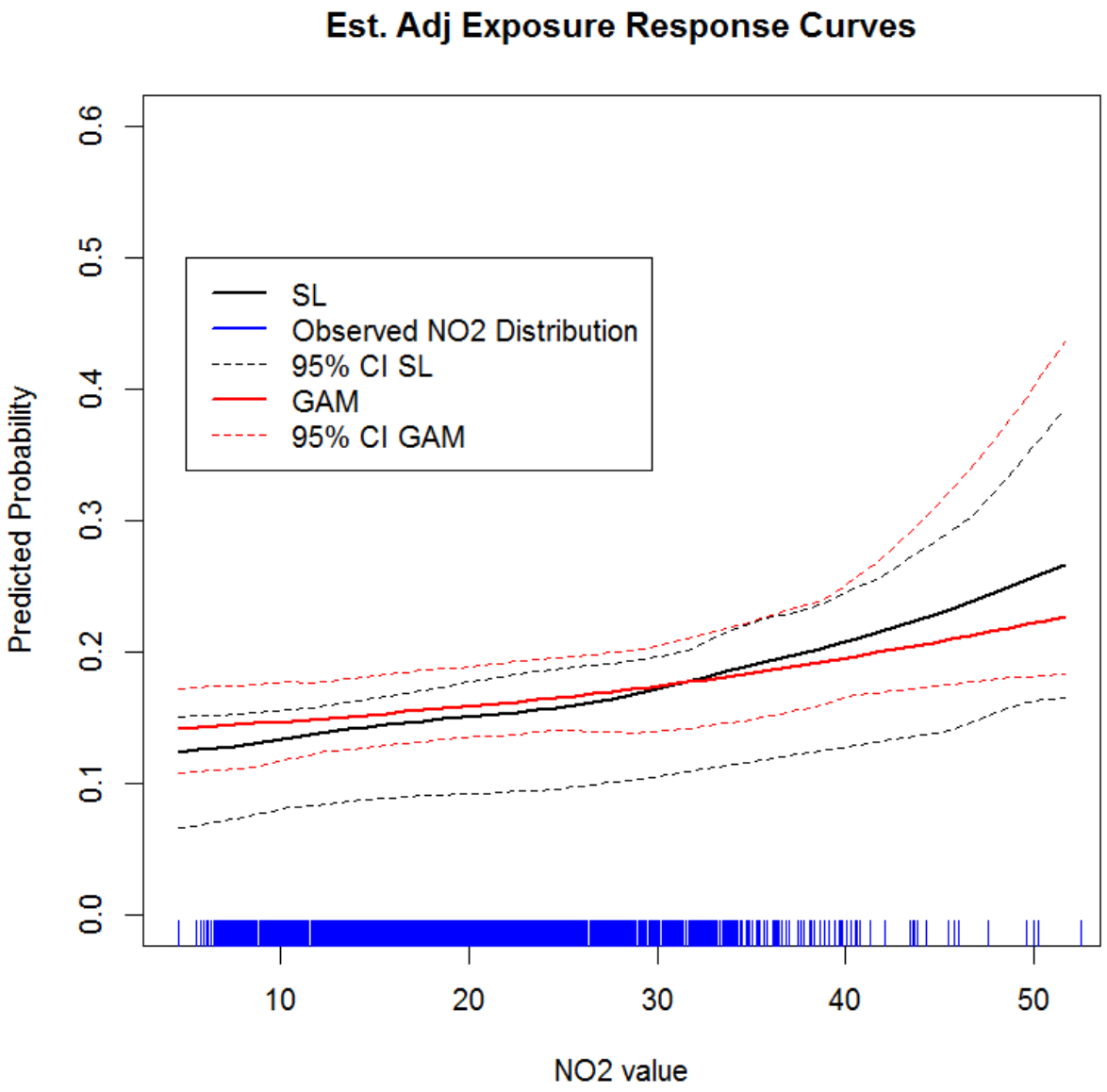

Figure 2: Estimated adjusted marginal exposure-response curve and CIs from SL and GAM. The range of the horizontal axis on this graph is based on the NO_2 distribution observed in the data. The blue bars on the horizontal axis indicate actual NO_2 levels that were observed. The vertical axis is the predicted probability of wheezing. The red lines are 95% confidence limits.

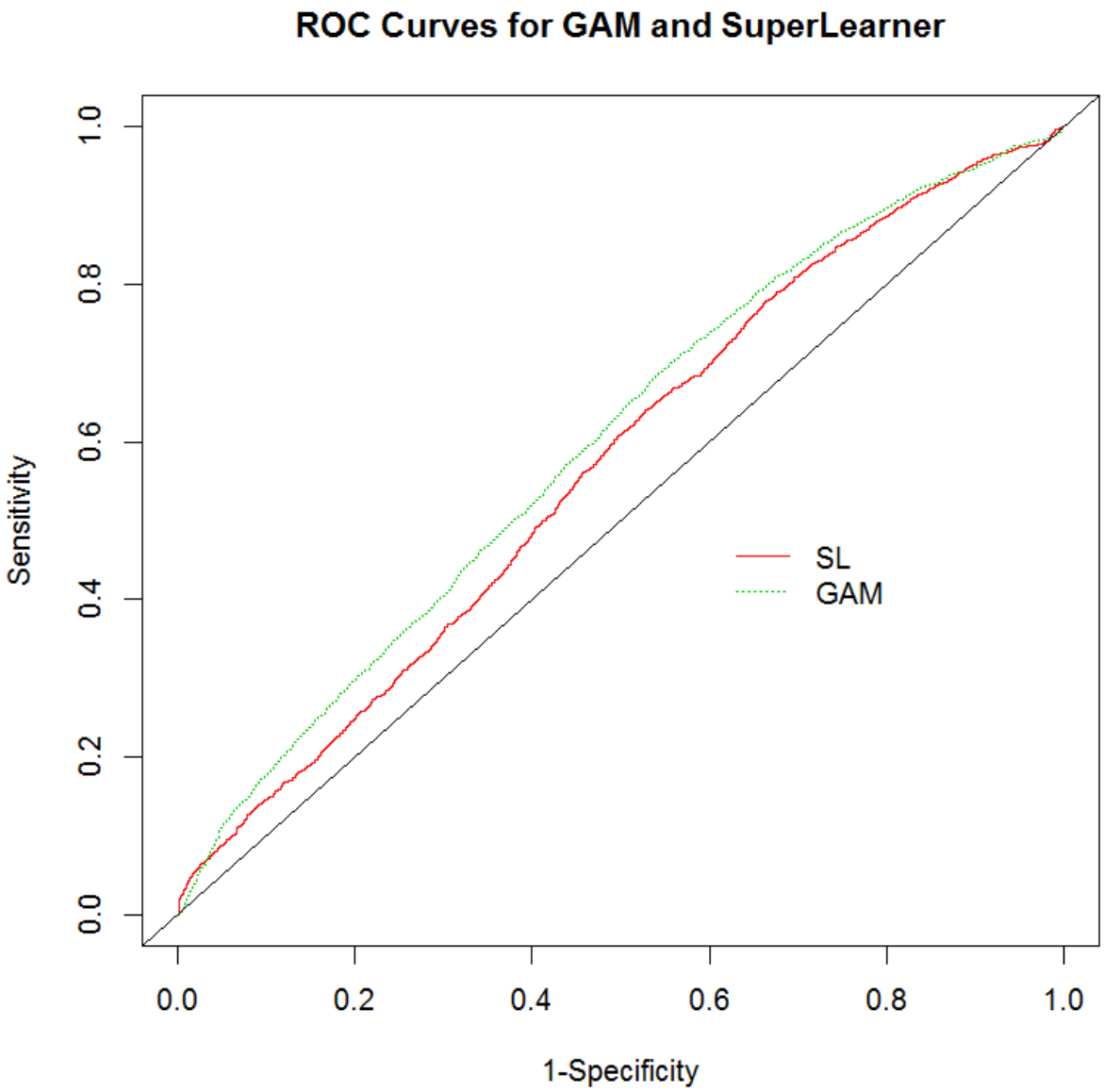

Figure 3: ROC curves of the cross-validated predictors based on GAM and SL. The black diagonal line indicates a predictor that does no better than chance.