NIH Public Access
Author Manuscript
Epidemiology. Author manuscript; available in PMC 2014 October 23.

Published in final edited form as:
Epidemiology. 2014 July ; 25(4): 583–590. doi:10.1097/EDE.0000000000000099.

Does exposure prediction bias health effect estimation? The relationship between confounding adjustment and exposure prediction

Matthew Cefalu and
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115, mcefalu@hsph.harvard.edu

Francesca Dominici
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, HSPH Building 2, 4th Floor, Boston, MA 02115

Abstract

In environmental epidemiology, we are often faced with two challenges. First, an exposure prediction model is needed to estimate the exposure to an agent of interest, ideally at the individual level. Second, when estimating the health-effect associated with the exposure, confounding adjustment is needed in the health-effects regression model. The current literature addresses these two challenges separately. That is, methods that account for measurement error in the predicted exposure often fail to acknowledge the possibility of confounding, while methods designed to control confounding often fail to acknowledge that the exposure has been predicted. In this paper, we consider exposure prediction and confounding adjustment in a health-effects regression model simultaneously. By using theoretical arguments and simulation studies, we show that the bias of a health-effect estimate is influenced by the exposure prediction model, the type of confounding adjustment used in the health-effects regression model, and the relationship between these two. Moreover, we argue that even with a health-effects regression model that properly adjusts for confounding, the use of a predicted exposure can bias the health-effect estimate unless all confounders included in the health-effects regression model are also included in the exposure prediction model. While these results of this paper were motivated by studies of environmental contaminants, they apply more broadly to any context where an exposure needs to be predicted.

Introduction

In the past two decades, there has been a wealth of epidemiological research on the health-effects of air pollution.[1–3] Studies have reported important associations between short-term and long-term exposure to ambient levels of air pollution and a wide range of adverse health outcomes.

Air pollution measurements are usually obtained from fixed monitoring locations, while data on health outcomes are generally available at the individual level with geocoded addresses

Correspondence to: Matthew Cefalu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

or as aggregated counts within a prespecified geographical region. The common approach to integrate these two types of data is to develop a statistical model for predicting levels of air pollution at places where the health outcomes are available.

Various methods can be employed to predict missing air pollution values, including nearest-neighbor and kriging approaches.[4,5] Recently, land-use regression has garnered much attention because of its ability to improve local variation in exposure prediction by incorporating land-use (geographic) covariates into the prediction model. Hoek et al[6] provides a review of land-use regression models, and others[7–14] have applied this methodology in epidemiological studies.

Another common issue in studies of air pollution and health is confounding,[15] which arises due to the complex dependencies that exist between air pollution, the health outcome of interest, and other covariates. Researchers employ expert knowledge in an attempt to control confounding through the use of covariates associated with both the exposure and the outcome. Great care is taken to minimize the magnitude of bias in the health-effect estimate, although it is unlikely that the bias can be completely removed. We use the term confounder here to define a covariate that is associated with the exposure, associated with the outcome independently of the exposure, and not on the causal pathway between the two.

Sheppard et al[15] provides a discussion of both confounding and exposure measurement error in air pollution epidemiology, and points out that exposure assessment should be evaluated in the context of health-effect estimation. With effect estimation in mind, it is known that: (1) better exposure prediction (i.e. smaller prediction error) does not necessarily lead to smaller mean squared error[16] of the health-effect estimate; and (2) confounding can lead to biased effect estimation.[17] However, the current literature treats confounding and exposure prediction as separate statistical issues. That is, methods that account for measurement error in the predicted exposure often fail to acknowledge the possibility of confounding, while methods designed to control confounding often fail to acknowledge that the exposure has been predicted.

We simultaneously consider exposure prediction and confounding adjustment in a health-effects regression model. Based on theoretical arguments, we show that using different sets of covariates in an exposure prediction model and in a health-effects regression model can bias the health-effect estimate. We provide a simulation study that illustrates this concept in the context of a cohort study on the association between long-term exposure to $PM_{2.5}$ and cardiovascular disease. We show that better prediction (higher R^2) does not always imply better effect estimation (smaller bias). Our results suggest that exposure prediction and confounding adjustment need to be considered simultaneously. In fact, we show that even under a correctly specified health-effects regression model that includes all confounders, the use of a predicted exposure can bias the health-effect estimate unless all of the confounders included in the health-effects regression model are also included in the exposure prediction model.

The concepts described in this paper, although motivated by epidemiological studies of air pollution and health, are broadly applicable to any context in which an exposure is predicted

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

with covariates that might also be confounders of the exposure-response relationship. In the discussion, we provide examples of the broader applicability of our results.

Confounding bias due to exposure prediction

Let $\mathbf{C}_i$ be a set of covariates for the ith observation, and assume that the outcome Y_i and the exposure X_i are generated under the following linear models:

$$Y_i = X_i\beta + \gamma_0 + \mathbf{C}_i\gamma + \varepsilon_i^y \quad (1)$$

$$X_i = \alpha_0 + \mathbf{C}_i\alpha + \varepsilon_i^x \quad (2)$$

where ε_i^y and ε_i^x are independent, normally distributed, mean-zero error terms with variances $\sigma^2_{y|xc}$ and $\sigma^2_{x|c}$. The true exposure X_i is assumed to be unobserved for all observations; therefore, an exposure prediction model is necessary.

Suppose interest lies in the estimation of the linear exposure-outcome relationship β, conditional on the covariates $\mathbf{C}_i$. Here, and throughout, no restriction is placed on γ or α, and individual components of the vectors are free to be 0.

The difficulty in estimating β in this setting is twofold: (1) in a practical application, we do not know the exact set of covariates **C** necessary for confounding adjustment; and (2) the exposure is not directly observed and must be predicted. Current literature fails to acknowledge these two important concepts simultaneously, and as such, researchers may be misguided when faced with these issues. Our results provide a starting point at understanding the complex relationship between exposure prediction and confounding adjustment and provide some basic guidance as to what can be expected in practice. We briefly present a few results here; for a fuller discussion and mathematical derivation, see the eAppendix.

First, imagine an admittedly unrealistic situation in which no confounding adjustment is needed ($\gamma_j = 0$ or $\alpha_j = 0$ for all $j \geq 1$). In such a situation, the bias of a health-effect estimate falls fully within the vast measurement error literature[18–21] and will therefore not be discussed here. Similarly, consider the case where the exposure is fully observed, but confounding adjustment is necessary ($\gamma_j \neq 0$ and $\alpha_j \neq 0$ for some $j \geq 1$). In this situation, one can rely exclusively on the large literature on confounding adjustment for standard regression modelling[22–24] and will therefore not be discussed here.

We address simultaneous consideration of these concepts — that is, when exposure prediction and confounding adjustment are both necessary, and when a potentially large set of covariates are available for confounding adjustment and exposure prediction. Using overlapping sets of covariates to predict exposure and adjust for confounding can lead to biased health-effect estimates. Intuitively, this can be expected; under this situation, the predicted exposure may be more correlated with the confounders than with the true exposure.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

When the goal of a study is health-effect estimation, the decision to include a potential confounder in either the exposure prediction model or the health-effects regression model needs to be based on more than just the predictive power of the potential confounder on the exposure or the strength of the relationship with the outcome. Rather, the decision needs to be based on some tradeoff between the two. To emphasize this concept, we will now discuss special cases where exposure prediction is either beneficial or detrimental to the goal of health-effect estimation under the common situation where confounding adjustment is necessary.

Confounding bias inflation due to exposure prediction under outcome model misspecification

Consider an oversimplified scenario in which all confounders are used to predict the exposure and no confounding adjustment is made in the health-effects regression model. Note that under the true models specified in Equations 1 and 2, this corresponds to a correct specification of the exposure prediction model but an incorrect specification of the health-effects regression model. While this would rarely occur in practice, it serves to illustrate that exposure prediction can increase the magnitude of confounding bias.

More specifically, assume that the true data-generating mechanism is given by Equations 1 and 2. We refer to confounding bias as the bias in the health-effect estimate from a health-effects regression model that fails to control for any confounding ($Y_i = \gamma_0 + X_i\beta + \varepsilon_i$). Let $W_i = \alpha_0 + \mathbf{C}_i\alpha$ be the predicted exposure with α_0 and α known from Equation 2. Consider fitting the health-effects regression model that uses the predicted exposure W_i in place of the true exposure X_i and fails to control for any confounding ($Y_i = \gamma_0 + W_i\beta + \varepsilon_i$). It can be shown that the bias of a health-effect estimate using the predicted exposure is always larger in magnitude than the confounding bias when using the true exposure. In this specific context, using a predicted exposure in a health-effects regression model will always lead to a more biased health-effect estimate than using the true exposure, and the degree to which the predicted exposure increases the bias is determined solely by the prediction accuracy of the exposure prediction model. Therefore, the bias in the health-effect estimate can be large even when the confounding bias is small if the correctly specified exposure prediction model has poor prediction power. The full analytical expression, with discussion, can be found in the eAppendix.

Such a result occurs because the correlation between the predicted exposure and the confounders that are missing from the health-effects regression model may be magnitudes larger than the correlation between the predicted exposure and the true exposure. Similar confounding bias inflation occurs under any health-effects regression model that fails to adequately adjust for confounding when using the true exposure prediction model.

As we will show in the subsequent sections, this relationship does not generalize to more complex settings. In fact, the bias of the health-effect estimate can either increase or decrease in magnitude if a subset of the confounders are used in the exposure prediction model (i.e. exposure prediction model misspecification). The bias depends on both the set of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

covariates used for exposure prediction and the set of covariates used for confounding adjustment in the health-effects regression model.

Confounding bias due to exposure prediction under exposure prediction model misspecification

In this section, we assume that confounding can be completely adjusted for in the health-effects regression model. We show that to return an unbiased health-effect estimate, either (1) the exposure prediction model must be correctly specified, (2) all confounders included in the health-effects regression model must be included in the exposure prediction model, or (3) the covariates used in the exposure prediction model must be uncorrelated with the confounders.

Assume that the data is generated under the following linear models:

$$Y_i = X_i\beta_x + \gamma_0 + \mathbf{C}_i^{(1)}\gamma_1 + \varepsilon_i^y \quad (3)$$

$$X_i = \alpha_0 + \mathbf{C}_i^{(1)}\alpha_1 + \mathbf{C}_i^{(2)}\alpha_2 + \mathbf{C}_i^{(3)}\alpha_3 + \varepsilon_i^x \quad (4)$$

where $\mathbf{C}^{(1)}$, $\mathbf{C}^{(2)}$, and $\mathbf{C}^{(3)}$ denote subsets of $\mathbf{C}$. Notice that in this data-generating scheme, there is only partial overlap in the sets of covariates in the two models, and that the necessary set of confounders is $\mathbf{C}^{(1)}$.

Assume that $\mathbf{C}^{(1)}$ is fully observed so that we can use a health-effects regression model that correctly adjusts for confounding by including $\mathbf{C}^{(1)}$ as a linear predictor. Under this correctly specified health-effects regression model, using a predicted exposure will bias the effect estimate unless (1) the exposure prediction model is correctly specified, (2) the covariates used in the exposure prediction are uncorrelated with the confounders $\mathbf{C}^{(1)}$, or (3) all of the confounders $\mathbf{C}^{(1)}$ are included in the exposure prediction model.

More specifically, assume that no data is collected on the set $\mathbf{C}^{(3)}$ and that $\alpha_3 \neq 0$, so that we are guaranteed misspecification of our exposure prediction model. Consider the health-effects regression model that correctly adjusts for confounding but uses a predicted exposure based only on $\mathbf{C}^{(2)}$ in place of the true exposure ($Y_i = W_i\beta + \gamma_0 + \mathbf{C}^{(1)}\gamma_1 + \varepsilon_i$). This exposure prediction model purposely excludes confounders $\mathbf{C}^{(1)}$. Under such a procedure, the health-effect estimate will be biased unless the covariates $\mathbf{C}^{(2)}$ used to predict the exposure are uncorrelated with the confounders $\mathbf{C}^{(1)}$ (result in eAppendix). A related result when estimating the health-effect in the presence of unmeasured confounding will be discussed below.

In most applications, the set of covariates used to predict the exposure will not be uncorrelated with the confounders, and excluding the confounders from the exposure prediction model will bias the health-effect estimate. However, an unbiased health-effect estimate can be obtained if the exposure prediction model is based on both $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$ (result in eAppendix). In other words, by including the confounders in the exposure

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

prediction model, the health-effects regression model that properly adjusts for confounding will remain unbiased. This result may seems counter to intuition at first glance, but is easily explained. By including the confounders $\mathbf{C}^{(1)}$ in the exposure prediction model, we guarantee that the variation in the predicted exposure that is used to estimate the health-effect is uncorrelated with the confounders. For a mathematical discussion, see the eAppendix.

In general, any misspecification in the health-effects regression model or the exposure prediction model will lead to a biased health-effect estimate. The magnitude of these biases will be difficult to compare under various levels of misspecification. More work needs to be done to determine the best strategy for minimizing the bias under model misspecification. However, these preliminary results suggest that under a health-effects regression model that properly adjusts for confounding, one must include confounders in the exposure prediction model to avoid bias.

Exposure prediction that removes confounding bias

In a simplified situation from the previous section, where the data is generated under Equations 3 and 4, assume that no data is collected on the set $\mathbf{C}^{(1)}$, so that direct control of confounding is impossible. Further, assume that $\alpha_3 = 0$ and that $\mathbf{C}^{(2)}$ is uncorrelated with the missing confounders $\mathbf{C}^{(1)}$. If the predicted exposure is based solely on these covariates $\mathbf{C}^{(2)}$ that are both predictive of the exposure and uncorrelated with the unmeasured confounding, then the resulting health-effect estimate will be unbiased (result in eAppendix). This result can be viewed as an instrumental variable approach where $\mathbf{C}^{(2)}$ is used as an instrument for X[25,26].

Such a result is applicable in the context of air pollution studies of health. Suppose that we believe there is spatial confounding in our study and are unable to control it because we do not have data at the desired level of spatial resolution. An unbiased health-effect estimate can be obtained by building an exposure prediction model that uses only those covariates that vary on a different spatial scale than the spatial confounding (or those covariates with little to no spatial structure). This is true because these covariates will be uncorrelated with the spatial confounding, and our previous result will hold.

Simulation setup and results in air pollution epidemiology

We have provided theoretical evidence that an exposure prediction model chosen solely on its ability to predict the true exposure may lead to a biased health-effect estimate. We now provide a simulated example that mimics a real cohort study of air pollution and health to show that better prediction (higher R^2) does not imply better effect estimation (smaller bias).

Consider a hypothetical cohort study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular disease in the New England region. Assume we have cardiovascular hospitalization rates over the study period for each of the 2165 zip codes in New England, and we wish to have available $PM_{2.5}$ levels for each of the 2165 zip codes. Of the 2165 zip codes, 57 have air pollution monitors within their boundaries, and the exposure for these zip codes can be measured directly as the mean monitor value during the study

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

period. For the remaining 2108 zip codes, we assume the exposure values are missing and need to be predicted.

Figure 1 provides a map of the 2165 zip codes in New England, with the 57 $PM_{2.5}$ monitoring locations marked with an **x**. We observe that the $PM_{2.5}$ monitors are sparse in New England, and tend to cluster near major population centers. As such, the spatial heterogeneity in $PM_{2.5}$ across New England will be difficult to capture based solely on spatial location (i.e. latitude and longitude).

The intention of this simulation is to illustrate how the choice of covariates used in the $PM_{2.5}$ prediction model will affect the estimated health-effect under a misspecified health-effects regression model that does not fully account for confounding. We generate 1000 realizations of our hypothetical cohort in the following manner:

1. Use the observed distribution of 9 land-use covariates for each zip code in New England. Table 1 provides a complete list and summary statistics for each land-use covariate considered.
2. Augment the 9 land-use covariates with one N (0,1) random variable, and denote the centered and standardized versions of these 10 covariates as $\mathbf{C}_i$.
3. Generate the exposure based on the relationship between the observed $PM_{2.5}$ levels and **C**. That is, fit the exposure model $X_i = \alpha_0 + \mathbf{C}_i\alpha + \varepsilon_x$ for the 57 zip codes that have observed $PM_{2.5}$ measurements, and use the resulting $\hat{\alpha_0} = 10.86$, $\hat{\alpha} = (-0.51, -0.44, -0.38, 0.35, 0.22, -0.16, 0.16, 0.13, -0.05, 0)$, and $\hat{\sigma}^2 = \hat{var}(\varepsilon_x)$ to generate a simulated "true" exposure as: $\tilde{X_i} = \hat{\alpha_0} + \mathbf{C}_i\hat{\alpha} + \mathcal{N}(0, \hat{\alpha^2})$
4. Generate the cardiovascular hospitalization rates using the regression model: $ln(Y_i) = \tilde{X_i}\beta + \mathbf{C}_i\gamma + \mathcal{N}(0, 0.467^2)$, where $\gamma = (0, 0.01, 0.11, -0.15, -0.13, 0.12, -0.14, -0.13, 0.06, 0.01)$ and $\beta = 0.04$. Other choices of γ were considered and are available in the eAppendix.
5. Remove the "true" $PM_{2.5}$ values $\tilde{X}$ from the dataset to reflect the zip code that are missing exposure. The final dataset contains 57 zip codes of (Y_i, $\tilde{X_i}$, $\mathbf{C}_i$) and 2108 zip codes of (Y_i, $\mathbf{C}_i$)

Table 2 summarizes the covariates that are included in each model of the data-generating mechanism. Note that the true exposure is generated using C_1 through C_9, while the true confounders are C_2 through C_9.

For ease of demonstration, this data generating mechanism purposely uses a nearly worst-case scenario; there is a large overlap in the set of covariates used to predict the exposure and those used to adjust for confounding. In reality, there will be partial overlap between these two sets. See the eAppendix for further discussion.

We will proceed using land-use regression to estimate $PM_{2.5}$ levels that are missing from the study. Once the land-use regression is used to estimate the missing $PM_{2.5}$ values, a health-effects regression is performed using a completed dataset that replaces the missing 2108 $PM_{2.5}$ values with their corresponding predicted values.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

The only remaining decision for the purpose of our simulation is which land-use covariates to include in the land-use regression model to predict the missing $PM_{2.5}$ values. Considering every combination of the land-use covariates would amount to $2^{10} = 1024$ possible models. Instead, we chose to consider 10 nested regression models that include the 10 land-use covariates in order of their true predictive power of $PM_{2.5}$. The following summarizes the steps used to predict $PM_{2.5}$ and estimate the resulting health-effect:

1. Fit the land-use regression model including only C_1 as a predictor for the 57 zip codes with observed $PM_{2.5}$
2. Estimate the 2108 missing $PM_{2.5}$ values, W, based on the model from Step 1
3. Estimate the effect of long-term $PM_{2.5}$ exposure on cardiovascular hospitalization rates using a regression model including only W as a predictor ($\ln(Y_i) = \gamma_0 + W_i\beta + \varepsilon_i$). The health-effects regression model is always misspecified.
4. Repeat Steps 1–3, but using $\{C_1, C_2\}$, $\{C_1, C_2, C_3\}$, …, $\{C_1, \ldots, C_{10}\}$ as predictors in the exposure regression model from Step 1. Each land-use regression model is misspecified, except for the one that includes $\{C_1, \ldots, C_9\}$

Note that in Step 3, we fit a regression model that fails to control confounding and gives a biased health-effect estimate. The magnitude of this bias, which is given in closed form in the eAppendix, is determined by a tradeoff between the bias due to lack of adjustment for confounding and the prediction accuracy of the $PM_{2.5}$ regression model. It does not depend on the true value of β. As such, we consider only one value of $\beta = 0.04$. However, if we had chosen to control confounding for some fixed set of covariates, results would be similar.

Figure 2 provides the R^2 from the land-use regression models and the corresponding bias of the health-effect estimate from the hypothetical study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular hospitalization rates. The land-use regression model that provides the health-effect estimate with the smallest bias is the one that includes the first five land-use covariates (% forest, % open space, % urban, traffic density, and elevation) and has corresponding R^2 value of less than 0.6. This occurs because the bias due to the lack of adjustment for confounding happens to be negated by the bias due to the measurement error in the exposure prediction. By including the two additional covariates distance to major road and point emissions, the R^2 can be increased to 0.7, but this results in a large bias. In this case, the measurement error induced by the exposure prediction no longer negates the confounding bias, and we are left with a biased health-effect estimate. Of the 10 models considered, 5 have a smaller bias than the model that uses the true exposure (the dotted line in Figure 2), suggesting that a predicted exposure can either improve or worsen effect estimation when compared to the true exposure in the presence of uncontrolled confounding.

This simple simulation illustrates that in the presence of uncontrolled confounding, a more accurate prediction of the exposure does not necessarily lead to a better health-effect estimate. Exposure prediction can exacerbate the problem of uncontrolled confounding, but all is not lost. Recall that in this hypothetical study, we purposefully fail to control for any confounding, but with a properly chosen $PM_{2.5}$ prediction model, we were able to return

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

nearly unbiased effect estimates. In that situation, the exposure prediction was beneficial for health-effect estimation, but we are able to determine this only because we know the true exposure and health-effect regression models.

In addition to fitting the data as described above, we illustrate that even under a health-effects regression model that properly adjusts for confounding, a predicted exposure can lead to a biased health-effect estimate. Note that C_2 through C_9 are the confounders. Thus, we proceed as follows:

1. Fit an incorrectly specified land-use regression model that includes only C_1 and C_{10} (and excludes C_2 through C_9) as predictors for the 57 zip codes with observed $PM_{2.5}$. We purposely included in the model only the covariates that are not confounders.
2. Estimate the 2108 missing $PM_{2.5}$ values, W, based on the model from Step 1.
3. Estimate the effect of long-term $PM_{2.5}$ exposure on cardiovascular hospitalization rates using a regression model including W and all of the necessary confounders C_2 through C_9 as predictors ($\ln(Y_i) = \beta W_i + \gamma_0 + \gamma_2 C_{2i} + \ldots + \gamma_9 C_{9i} + \varepsilon_i$). Note that using the true exposure in this model would yield an unbiased health-effect estimate.

Using the above data-fitting algorithm, the bias of the health-effect estimate is 0.01, corresponding to a 25% bias. This occurs because, although confounding bias is zero, the exposure model has been misspecified. To avoid this bias, the exposure prediction model that includes all confounders in addition to the other covariates can be used in Step 1. Doing so makes the bias approximately 0 (the simulated value is 0.0002). This verifies a few of our theoretical results previously discussed. First, even under a health-effects regression model that properly adjusts for confounding, a predicted exposure can bias the health-effects estimate. Second, including all confounders in the exposure prediction model makes the estimated health-effect unbiased.

Discussion

We simultaneously consider two of the most important and challenging issues in environmental epidemiology: exposure prediction and confounding adjustment. Although the motivation and terminology come from air pollution epidemiology, results apply to any context in which exposure measurements are incomplete and must be predicted. Examples include exposures to herbicides or pesticides, burn pits in Iraq and Afghanistan, low-level radio waves, dietary intake, and blood lead levels[27–32]. In each, the exposure of interest is not completely observed and must be predicted, while confounding adjustment is necessary when fitting the health-effects regression model.

Current statistical methods dealing with missing exposure and confounding adjustment treat the two topics as distinct. For example, methods to overcome missing exposure rely on exposure prediction, and exposure prediction can be viewed as a measurement error problem.[18–21] Methods for exposure prediction are concerned only with bias of the health-effect estimates due to measurement error associated with the prediction of the true

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

exposure; these methods do not consider how predicting exposure with covariates that are correlated with the outcome might bias the health-effect estimates. Similarly, methods designed for confounding adjustment do not acknowledge that the exposure has been predicted. For example, the approach of Wang et al[33] was designed for the selection of confounders in the context of linear models for both the outcome and the exposure when the exposure has been fully observed.

New statistical methods are needed to simultaneously predict exposure while adjusting for confounding. While the results of our paper indicate that under a properly specified health-effects regression model, all confounders should be used to predict the exposure, the optimal strategy is less clear when there is uncertainty surrounding the covariates that should be included in both the exposure prediction model and the health-effects regression model. The decision to include a covariate in the outcome or the exposure model needs to be based on both the predictive power of the covariate on the exposure and the strength of the relationship with the outcome. An extension of Wang et al[33] into the context of missing exposure could provide a foundation of methodologies used to simultaneously predict exposure and control confounding.

We framed our results with the priority of returning an unbiased health-effect estimate, and we ignored the possibility of a bias-variance tradeoff. We do not include any results on the variance in order not to distract from the main points. However, we do not believe that there is a simple bias-variance tradeoff when simultaneously considering exposure prediction and confounding adjustment.

Our results do not address how spatial smoothing will affect the bias of a health-effect in the presence of unmeasured spatial confounding. However, it is reasonable to postulate similar results on the bias of a health-effect estimate when employing spatial smoothing. Such results would be related to the work of Dominici et al[34]; these researchers provide results to reduce confounding bias in the pollution-mortality relationship due to unmeasured time-varying factors such as season and inuenza epidemics in the context of time-series studies. One could adapt their results for cross-sectional studies of air pollution and health by indexing by space instead of time.

Issues of bias have been presented in the context of cross-sectional studies. There is a likely statistical parallel for time-series studies. If missing exposure values are imputed using covariates that are temporally correlated with both the exposure and the outcome, then similar biases are likely. For example, meteorological covariates could be temporally correlated with both air pollution and health.

We assumed simple linear relationships between the outcome, the exposure, and the confounders. Deriving analytic results for biases under more complex models is impractical; however, similar biases are expected. Greater care is needed when using predicted exposure values in epidemiological studies of health.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

References

1. Dominici F, Sheppard L, Clyde M. Health Effects of Air Pollution: A Statistical Review. International Statistical Review. 2003; 71:243276.
2. Pope CA. Mortality Effects of Longer Term Exposures to Fine Particulate Air Pollution: Review of Recent Epidemiological Evidence. Inhalation Toxicology. 2007; 19:3338.
3. Breysse P, Delfino R, Dominici F, Elder A, Frampton M, Froines J, Geyh A, Godleski J, Gold D, Hopke P, Koutrakis P, Li, Ning Oberdrster G, Pinkerton K, Samet J, Utell M, Wexler A. US EPA particulate matter research centers: summary of research results for 2005–2011. Air Qual Atmos Health. 2012
4. Oliver MA, Webster R. Kriging: a method of interpolation for geographical information systems. International Journal of Geographical Information System. 1990; 4:313332.
5. Madsen L, Ruppert D, Altman N. Regression with spatially misaligned data. Environmetrics. 2008; 19:453467.
6. Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, Briggs D. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmospheric environment. 2008; 42:75617578.
7. Henderson SB, Beckerman B, Jerrett M, Brauer M. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. Environmental science & technology. 2007; 41:24222428.
8. Ross Z, Jerrett M, Ito K, Tempalski B, Thurston GD. A land use regression for predicting fine particulate matter concentrations in the New York City region. Atmospheric Environment. 2007; 41:2255–2269.
9. Yanosky JD, Paciorek CJ, Schwartz J, Laden F, Puett R, Suh HH. Spatio-temporal modeling of chronic PM10 exposure for the Nurses Health Study. Atmospheric Environment. 2008; 42:4047–4062. [PubMed: 19584946]
10. Sahsuvaroglu T, Jerrett M, Sears MR, McConnell R, Finkelstein N, Arain A, Newbold B, Burnett R, et al. Spatial analysis of air pollution and childhood asthma in Hamilton, Canada: comparing exposure methods in sensitive subgroups. Environmental Health. 2009; 8:14. [PubMed: 19338672]
11. Neupane B, Jerrett M, Burnett RT, Marrie T, Arain A, Loeb M. Long-term exposure to ambient air pollution and risk of hospitalization with community-acquired pneumonia in older adults. American journal of respiratory and critical care medicine. 2010; 181:4753.
12. Kloog I, Coull BA, Zanobetti A, Koutrakis P, Schwartz JD. Acute and Chronic Effects of Particles on Hospital Admissions in New-England. PloS one. 2012a; 7:e34664. [PubMed: 22529923]
13. Kloog I, Melly S, Ridgway W, Coull B, Schwartz J. Using new satellite based exposure methods to study the association between pregnancy pm2.5 exposure, premature birth and birth weight in Massachusetts. Environmental Health. 2012b; 11:40. [PubMed: 22709681]
14. Cesaroni G, Badaloni C, Gariazzo C, Stafoggia M, Sozzi R, Davoli M, Forastiere F, et al. Long-Term Exposure to Urban Air Pollution and Mortality in a Cohort of More than A Million Adults in Rome. Environmental health perspectives. 2013
15. Sheppard L, Burnett R, Szpiro A, Kim S, Jerrett M, Pope C, Brunekreef B. Confounding and exposure measurement error in air pollution epidemiology, Air Quality. Atmosphere & Health. 2011; 114
16. Szpiro A, Paciorek C, Sheppard L. Does more accurate exposure prediction necessarily improve health effect estimates? Epidemiology. 2011a; 22:680685.
17. Pope CA III, Burnett RT. Confounding in air pollution epidemiology: the broader context. Epidemiology. 2007; 18:424426.
18. Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics. 2009; 10:258274.
19. Szpiro AA, Sheppard L, Lumley T. Efficient measurement error correction with spatially misaligned data. Biostatistics. 2011b; 12:610623.
20. Carroll, RJ.; Ruppert, D.; Stefanski, LA. Measurement Error in Nonlinear Models. Chapman and Hall/CRC; 1995.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

21. Basagana X, Aguilera I, Rivera M, Agis D, Foraster M, Marrugat J, Elosua R, Kunzli N. Measurement Error in Epidemiologic Studies of Air Pollution Based on Land-Use Regression Models. American Journal of Epidemiology. 2013; 178:1342–1346. [PubMed: 24105967]
22. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine. 2004; 23:29372960.
23. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11:550560.
24. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:4155.
25. Pearl J. On a class of bias-amplifying variables that endanger effect estimates. arXiv preprint arXiv. 2012; 1203.3503
26. Bhattacharya J, Vogt WB. Do Instrumental Variables Belong in Propensity Scores? International Journal of Statistics & Economics. 2012; 9:107127.
27. Choudhury H, Harvey T, Thayer WC, Lockwood TF, Stiteler WM, Goodrum PE, Hassett JM, Diamond GL. Urinary cadmium elimination as a biomarker of exposure for evaluating a cadmium dietary exposure biokinetics model. Journal of Toxicology and Environmental Health Part A. 2001; 63:321–350. [PubMed: 11471865]
28. Frei P, Mohler E, Burgi A, Frohlich J, Neubauer G, Braun-Fahrlander C, Roosli M. A prediction model for personal radio frequency electromagnetic ffeld exposure. Science of the total environment. 2009; 408:102108.
29. Lewin MD, Sarasua S, Jones PA. A multivariate linear regression model for predicting childrens blood lead levels based on soil lead levels: A study at four superfund sites. Environmental research. 1999; 81:5261.
30. Institute of Medicine (US) Committee on the Long-Term Health Consequences of Exposure to Burn Pits in Iraq and Afghanistan. Longterm Health Consequences of Exposure to Burn Pits in Iraq and Afghanistan. National Academies Press; 2011.
31. Committee on Making Best Use of the Agent Orange Exposure Reconstruction Model. The Utility of Proximity-Based Herbicide Exposure Assessment in Epidemiologic Studies of Vietnam Veterans. The National Academies Press; 2008.
32. National Research Council (US) Committee to Assess Potential Health Effects from Exposures to PAVE PAWS Low-Level Phased-Array Radiofrequency Energy. An Assessment of Potential Health Effects from Exposure to PAVE PAWS Low-level Phased-array Radiofrequency Energy: Letter Report. National Academies; 2006. [March 7, 2006]
33. Wang C, Parmigiani G, Dominici F. Bayesian Effect Estimation Accounting for Adjustment Uncertainty. Biometrics. 2012
34. Dominici F, McDermott A, Hastie TJ. Improved semiparametric time series models of air pollution and mortality. Journal of the American Statistical Association. 2004; 99:938948.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

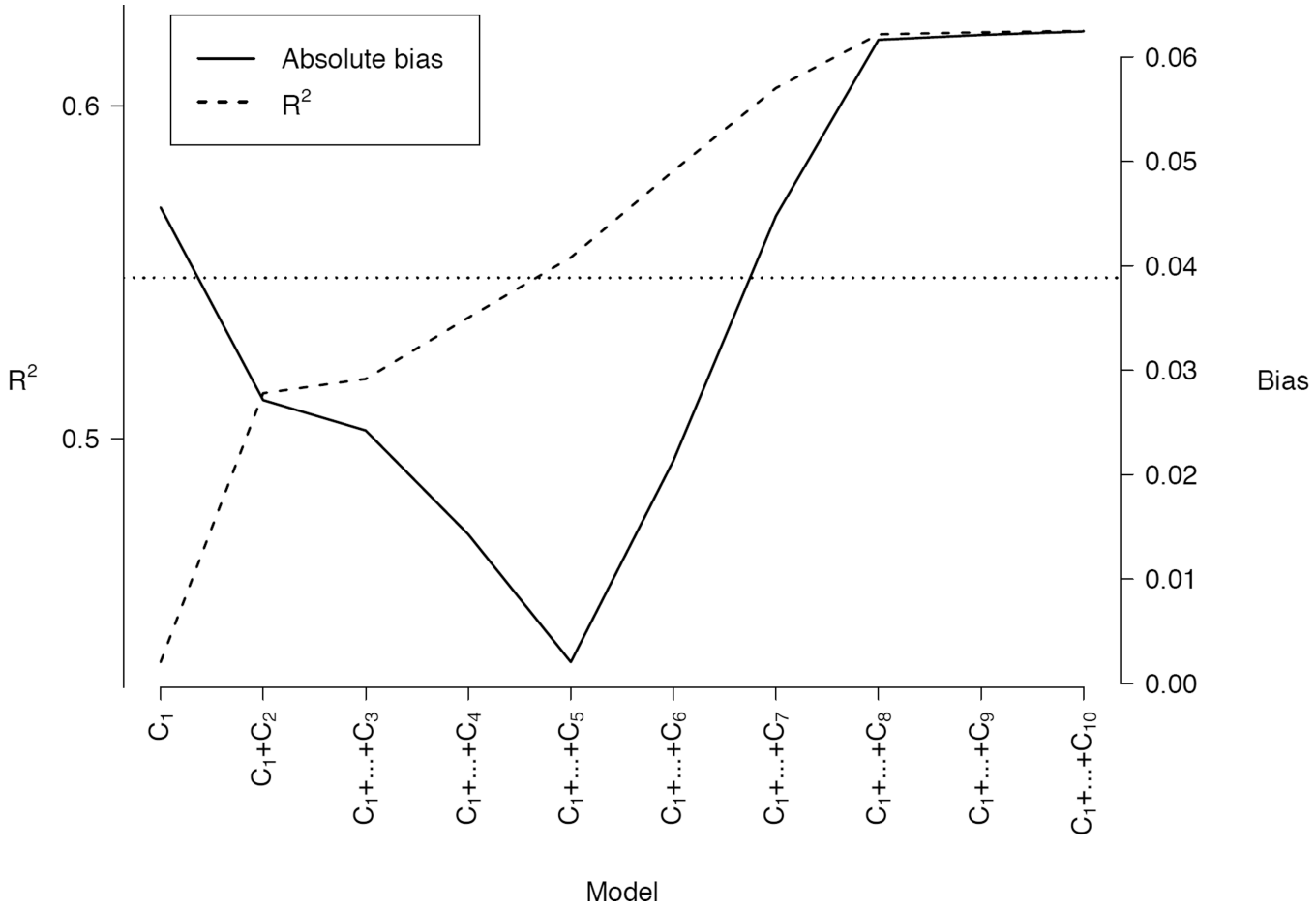

Figure 1.

Map of the 2165 zip codes in New England with the 57 $PM_{2.5}$ monitoring locations marked with an **x**

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

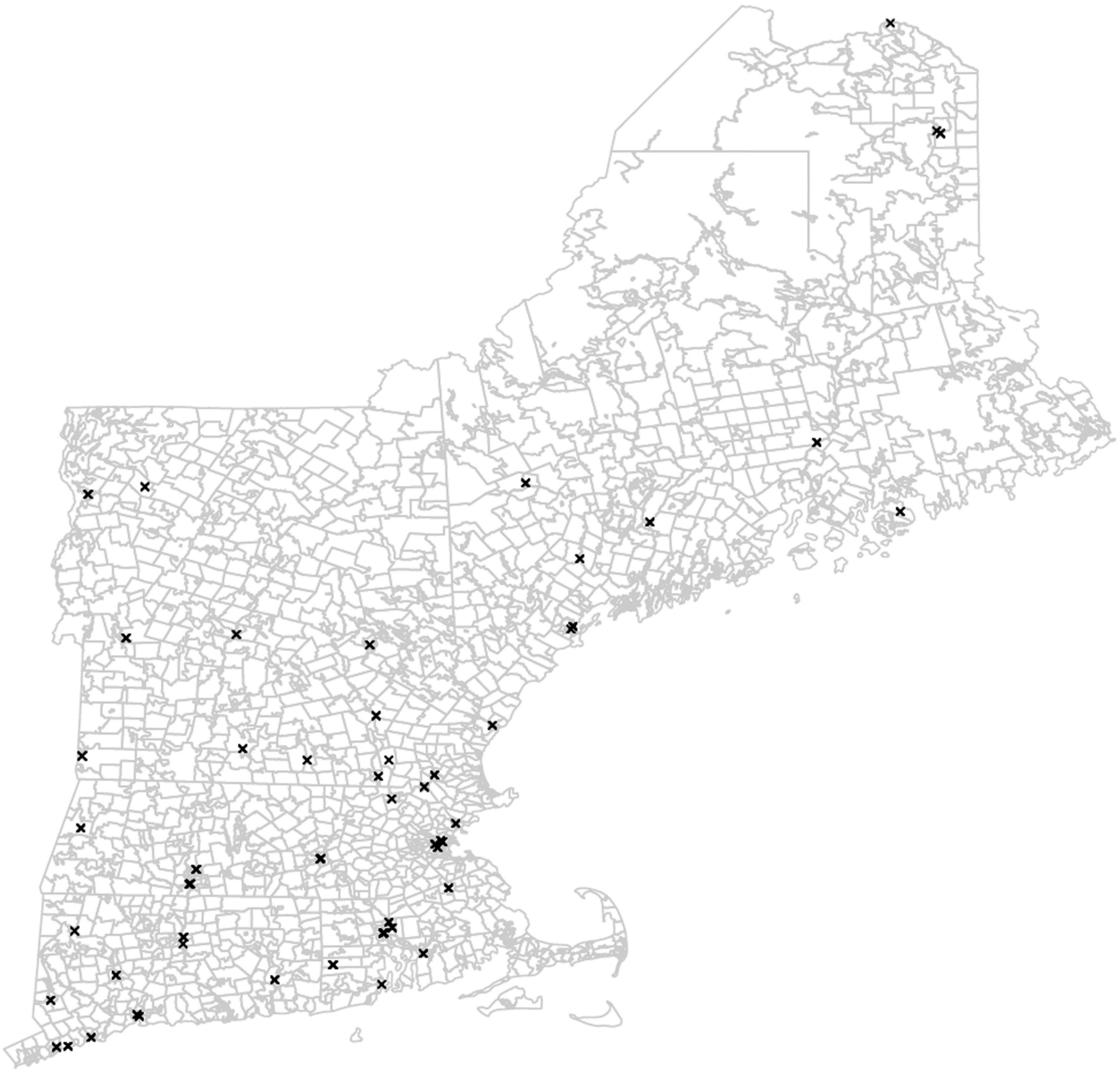

Figure 2.

The R^2 of the land-use regression models and the corresponding bias of the health-effect estimate from the hypothetical study of the association between long-term exposure to $PM_{2.5}$ and cardiovascular hospitalization rates in the New England region. The solid line is the bias, the dashed line the R^2, and the dotted line corresponds to the bias due to the lack of adjustment for confounding.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Summary of 9 Land-use Covariates in New England

Covariate	Minimum	1st quartile	Median	3rd quartile	Maximum
% Forest	0.36	31.54	55.99	71.56	96.19
% Open space	2.71	64.47	89.07	94.91	100
% Urban	0	1.32	4.71	22.58	92.50
Traffic density	0	13.84	20.97	36.82	122.16
Elevation	10	40	120	240	1160
Distance to major road	0	0	2.91	13.22	133.87
Point emissions	0.001	0.001	0.0491	1.23	2015.46
Area emissions	0.003	0.260	1.19	8.16	86.01
Population per sq. km	0.13	25.71	105.18	363.71	4044.88

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Summary of Covariates Included in Each Model of the Simulation Study

Covariate	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
Exposure model	X	X	X	X	X	X	X	X	X	
Outcome model		X	X	X	X	X	X	X	X	X
Confounders		X	X	X	X	X	X	X	X	