Does More Accurate Exposure Prediction Necessarily Improve Health Effect Estimates?

Adam A. Szpiro,[a] *Christopher J. Paciorek,*[b,c] *and Lianne Sheppard*[a,d]

Abstract: A unique challenge in air pollution cohort studies and similar applications in environmental epidemiology is that exposure is not measured directly at subjects' locations. Instead, pollution data from monitoring stations at some distance from the study subjects are used to predict exposures, and these predicted exposures are used to estimate the health effect parameter of interest. It is usually assumed that minimizing the error in predicting the true exposure will improve health effect estimation. We show in a simulation study that this is not always the case. We interpret our results in light of recently developed statistical theory for measurement error, and we discuss implications for the design and analysis of epidemiologic research.

(*Epidemiology* 2011;22: 680–685)

Downloaded from https://journals.lww.com/epidem by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3ZPR5XgVoeNmmO9+vnCR8d0gZFXbN9kcRk23CwoS/9DgQX52Y55w= on 02/19/2019

There has been a major effort in air pollution epidemiology research to develop statistical models to predict exposures at subjects' locations in situations where measurements at the desired locations are not available.[1–7] These efforts assume that exposure predictions with less measurement error relative to the unknown true values will improve health effect estimation.[8–10] We demonstrate, in a simulation study, that this assumption is not always true, and we interpret our results using recently developed statistical theory for measurement error resulting from spatially misaligned data.[11]

MATHEMATICAL FRAMEWORK AND SIMULATION STUDY

Most modern statistical models for predicting long-term average air pollution concentrations are based on land-use regression. In land-use regression modeling, a linear regression model with geographic (land-use) covariates such as population density, proximity to traffic, and proximity to commercial areas is fit to monitoring data and then used to predict concentrations at subjects' locations. Elaborations on this framework account for spatial and spatiotemporal correlation and various approaches to model selection, but land-use regression remains a central component. We focus on a pure land-use regression model in this paper.

Stochastic Data-generating Model

Consider an association study with the $N \times 1$ vector of observed health outcomes Y, $N \times 1$ vector of exposures X, and $N \times m$ matrix of covariates Z. Assume a linear regression model

$$Y = \beta_0 + X\beta_X + Z\beta_Z + \varepsilon, \qquad (1)$$

with coefficient of interest β_X and ε an $N \times 1$ random vector with independent elements distributed as Gaussian random variables with mean 0 and variance σ_ε^2 (ie, $N(0,\sigma_\varepsilon^2)$).

We are interested in the situation where Y and Z are observed, but instead of X, we observe the $N^* \times 1$ vector X^* of exposures at various locations. N^* is the number of exposure monitors. Assume that X, the subjects' exposures, and X^*, the exposure concentrations measured at the monitors, are jointly distributed as

$$\begin{pmatrix} X \\ X^* \end{pmatrix} = \begin{pmatrix} S \\ S^* \end{pmatrix}\alpha + \begin{pmatrix} \eta \\ \eta^* \end{pmatrix}. \qquad (2)$$

In this expression, S and S^* are random $N \times k$ and $N^* \times k$ dimensional matrices of the k geographic covariates used in the land-use regression model observed without error, α is an unknown $k \times 1$ vector of coefficients, and η and η^* are independent vectors with elements distributed as $N(0,\sigma_\eta^2)$. The stochasticity in S and S^* derives from random selection of subject and monitor locations. If the exposure model is known, it is standard practice to estimate α based on X^* and then use $W = S\hat{\alpha}$ in place of X in equation (1) to estimate β_X. That is, predictions from the land-use regression model are used as estimated exposures in place of the unknown true values, a form of regression calibration.[12]

Submitted October 2010; accepted 7 April 2011; posted 30 June 2011.

From the [a]Department of Biostatistics, University of Washington, Seattle, WA; [b]Department of Biostatistics, Harvard School of Public Health, Boston, MA; [c]Department of Statistics, University of California, Berkeley, CA; and [d]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA.

Supported by the United States Environmental Protection Agency through R831697 and Assistance Agreement CR-83407101 and by the National Institute of Environmental Health Sciences through R01-ES009411 and 5P50ES015915 (Drs. Szpiro and Sheppard) and by the National Institute of Environmental Health Sciences through R01 ES017017 (Dr. Paciorek).

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Adam A. Szpiro, Department of Biostatistics, University of Washington, Seattle, WA 98195. E-mail: aszpiro@u.washington.edu.

Copyright © 2011 by Lippincott Williams & Wilkins
ISSN: 1044-3983/11/2205-0680
DOI: 10.1097/EDE.0b013e3182254cc6

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

We quantify the accuracy in approximating X by W by

$$R^2_W = 1 - \sum_{i=1}^{N} \frac{(W_i - X_i)^2}{\sum_{j=1}^{N}\left(X_j - \frac{1}{N}\sum_{k=1}^{N} X_k\right)^{2'}} \qquad (3)$$

where larger R^2_W values correspond to less measurement error. This defines an out-of-sample measure of prediction accuracy, as it is based on prediction error at subjects' locations, and it is not subject to bias from overfitting the exposure model to the monitoring data.[13] R^2_W is a random quantity that varies for each realization of the data-generating model, and we denote its expectation $\bar{R}^2_W$.

There are a number of criteria for evaluating the validity and reliability of health effect estimates. We consider bias, standard deviation, root mean squared error, and coverage probability (the proportion of 95% confidence intervals that include the true β_X).[14]

Misspecified Exposure Model

We generally do not know the exact form of the exposure model and may use a misspecified model for prediction. One form of model misspecification is to omit a geographic covariate from the land-use regression model. This corresponds to observing only the $N \times (k-1)$ and $N^* \times (k-1)$ matrices S' and $S^{*'}$ obtained by deleting the kth columns of S and S^*. We then estimate the corresponding $(k-1) \times 1$ vector of coefficients α' and replace X in equation (1) by $W' = S'\hat{\alpha}'$ to obtain $\hat{\beta}'_X$. We denote measures of exposure prediction accuracy $R^{2'}_W$ and $\bar{R}^{2'}_W$ as in the case of the correctly specified exposure model.

We generally expect R^2_W to be larger than $R^{2'}_W$, which from the perspective of exposure modeling implies that the correctly specified exposure model gives better predictions than the misspecified one. It is reasonable to expect that this will also lead to improved health effect estimation. However, in the next subsection, we will demonstrate a class of examples in which R^2_W is consistently larger than $R^{2'}_W$, but $\hat{\beta}'_X$ has more error than $\hat{\beta}_X$ as measured in terms of bias, variance, root mean squared error, and coverage probability. We emphasize that R^2_W is not inflated by overfitting, because it is based on the correctly specified exposure model and quantifies out-of-sample prediction accuracy at subjects' locations.

Simulation Study

We set $k = 4$ (3 geographic covariates and an intercept) and consider scenarios with N between 100 and 10,000 subjects and $N^* = 100$ monitors. We assume the 3 geographic covariates are independent of each other at all locations and are independent between subjects. In particular, for each subject i, we assume the jth geographic covariate S_{ij} is independently distributed as $N(0,1)$.

Similarly, we assume the S^*_{ij} are distributed as $N(0,1)$ for $j = 1,2$, but the third geographic covariate for the monitoring sites is distributed as $N(0,\sigma^2)$ for $\sigma^2 = 0.1$, 1.0, or 4.0. Finally, we set $\alpha_0 = 0$, $\alpha_j = 4$ for $j = 1, 2, 3$, $\beta_0 = 1$, $\beta_X = 2$, $\sigma_\varepsilon = 25$, and $\sigma_\eta = 4$, and we assume there are no additional covariates Z. Example simulation code in R[15] can be found in the eAppendix (http://links.lww.com/EDE/A497).

The choice of σ^2 controls the level of variability in the third geographic covariate at the monitoring locations. By comparing the misspecified model (ie, the model that does not contain the third geographic covariate) to the correctly specified full model, we are able to assess the added value of including the third geographic covariate in predictions, depending on its variability. The situation with $\sigma^2 = 0.1$ is of particular interest, as it represents a geographic covariate that has limited variability in the monitoring data compared with the other geographic covariates but is equally variable in the subject data where it will be used to predict exposures. This is realistic, for example, if the covariate measures near-road traffic exposure. Regulatory monitors are often sited away from roadways to measure background pollution levels, and so they may not span the full range of covariate values relevant for predicting exposures at subjects' home locations—a significant fraction of which are near major roads.

In the Table and Figure 1, we show the results from 80,000 Monte Carlo simulations with $N = 10{,}000$ subjects, $N^* = 100$ monitoring sites, and $\sigma^2 = 1.0$. The coefficient for the third geographic covariate α_3 is estimated well in the full model and is statistically significant in all simulations. The corresponding exposure prediction accuracy R^2_W is consistently near 0.75, compared with $R^{2'}_W$ near 0.50 with the misspecified model. Health effect estimation efficiency is improved by using the correctly specified exposure model, which gives a standard deviation for $\hat{\beta}_X$ of 0.12 compared with 0.21 for $\hat{\beta}'_X$ with the misspecified model. The coverage probabilities for both models are poor, as the standard error estimates fail to account for exposure measurement error. The correctly specified exposure model results in a modest improvement in coverage probability, although it also introduces slightly more bias than the misspecified model.

Analogous results are shown in the Table and Figure 2 for $\sigma^2 = 0.1$, representing a situation where one of the geographic covariates is less variable in the distribution of monitoring locations than are the other geographic covariates. The smaller value of σ^2 results in more variability in estimating α_3, but this parameter is still estimated well and is statistically significant in 83% of Monte Carlo simulations. There is clear improvement in the exposure predictions from using the full model with R^2_W at least 0.67 in 95% of simulations, as compared with the misspecified model with $R^{2'}_W$ consistently near 0.50. But in this situation, the health effect estimates are more precise when we use the misspecified exposure model, with the standard deviation of $\hat{\beta}'_X$ equal to

© 2011 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

TABLE. Results From 80,000 Monte Carlo Simulations With $N = 10{,}000$ and $N^* = 100$

	$\sigma^2 = 1$		$\sigma^2 = 0.1$	
	Correct Model	**Misspecified Model**	**Correct Model**	**Misspecified Model**
Exposure predictions				
$\bar{R}^2_W$	0.74	0.49	0.73	0.50
$\overline{\mathrm{Var}}(W)$	48.5	32.7	50.2	32.3
Exposure model parameter estimate $\hat{\alpha}_3$				
Standard deviation	0.41	—	1.37	—
Statistically significant ($P < 0.05$)	100%	—	83%	—
Health effect parameter estimate $\hat{\beta}_X$				
Bias	−0.007	−0.001	−0.035	0.001
Standard deviation	0.12	0.21	0.23	0.16
RMSE	0.12	0.21	0.23	0.16
E(SE)	0.038	0.049	0.038	0.049
95% CI coverage	45%	35%	26%	46%

The $\bar{R}^2_W$ and $\overline{\mathrm{Var}}(W)$ are the out-of-sample prediction R^2_W and variance of predicted exposures, respectively, averaged over 80,000 Monte Carlo simulations. The standard deviation and fraction of Monte Carlo runs statistically significant are given for estimates of $\hat{\alpha}_3$ in the correctly specified exposure model only, because this parameter is not included in the misspecified model. The bias, standard deviation, root mean squared error (RMSE), and 95% confidence interval (CI) coverage are given for estimates of the health effect parameter $\hat{\beta}_X$. We also report the average estimated standard error (SE) for $\hat{\beta}_X$. The Monte Carlo standard error in estimating the bias of $\hat{\beta}_X$ in all the models is less than 0.001. (R code to produce these results is provided in an electronic appendix, http://links.lww.com/EDE/A497.)

0.16, compared with 0.23 for $\hat{\beta}_X$ using the fully specified model. The misspecified model also results in less bias and a modest improvement in coverage probability.

We vary the number of subjects as well as σ^2 and summarize the results in Figure 3 by plotting the difference between the standard deviation of $\hat{\beta}'_X$, based on the misspecified exposure model, and $\hat{\beta}_X$, based on the correct exposure model, on the vertical axis against N on the horizontal axis; a positive difference indicates that the correctly specified model is more efficient. We restrict to 5000 Monte Carlo simulations because this is sufficient to estimate the standard deviations (the biases are smaller and require more Monte Carlo simulations). The difference is positive for $\sigma^2 = 1.0$ and 4.0, consistent with the prior expectation that more accurate exposure predictions result in more efficient health effect estimation. But it is negative for $\sigma^2 = 0.1$ except for the case where there are only $N = 100$ subjects, demonstrating that in larger health studies the misspecified exposure model results in more efficient health effect estimation, even though it gives less accurate exposure predictions. For all simulations, the average out-of-sample exposure model prediction accuracies are $\bar{R}^2_W$ between 0.73 and 0.75 for the correctly specified model and $\bar{R}^{2'}_W$ between 0.49 and 0.50 for the misspecified model that omits the third geographic covariate.

THEORETICAL INTERPRETATION IN A MEASUREMENT ERROR FRAMEWORK

The results of our simulation study seem paradoxical in that more accurate exposure predictions do not necessarily lead to improved health effect estimation. The Table shows that for $\sigma^2 = 0.1$, the correctly specified model consistently gives more variable exposure predictions and more accurate out-of-sample prediction, compared with the misspecified exposure model. However, a small part of the additional exposure variability is induced by error in estimating α_3, which leads to less efficient estimation of β_X. These findings can be understood in a theoretical context by referring to the statistical measurement error framework developed for this setting.[11,12]

Briefly, for a fairly general class of exposure models there are 2 components to the measurement error. The Berkson-like component of error results from smoothing the exposure surface using a model that may not account for all sources of variation and can be thought of as the part of the true exposure that is not predictable from the model. It is similar to standard Berkson error[16] in that it inflates the standard deviation of the health effect estimate and introduces little or no bias. However, it is different from Berkson error in that it is correlated in space and is not completely independent of the predicted exposures.[11,12] The classical-like component comes from uncertainty in estimating the exposure model parameters. It is similar to classical measurement error in that it is a source of variability in the predicted exposures and can introduce bias in health effect estimates as well as change their standard errors. The classical-like component is also different from classical measurement error in that the additional variability from exposure model parameter estimation is shared across all prediction locations rather than being independent.[11]

For the simple land-use regression exposure model considered here, the Berkson-like component is pure Berkson error because there is no spatial dependence structure in η and η^* and the S_{ij} and S^*_{ij} are independent. When we use the correctly specified exposure model, the Berkson error is just η, but misspecifying the model by omitting the third geographic covariate increases the Berkson error substantially, resulting in a degradation of prediction accuracy. However, Berkson error plays the same role mathematically as the random ε in the disease model, and so its impact on the health effect estimation error diminishes for large N. On the other hand, each coefficient that needs to be estimated in the exposure model contributes to the classical-like error, and this part of the error remains important regardless of the number of subjects. In some situations, this could result in a bias-variance tradeoff

© 2011 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

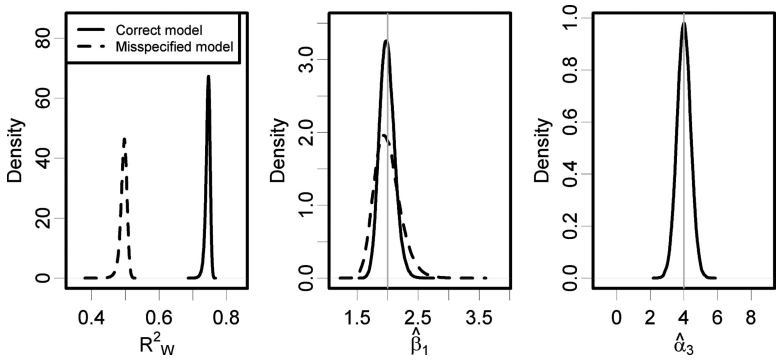

FIGURE 1. Results from 80,000 Monte Carlo simulations with $N = 10{,}000$, $N^* = 100$, and $\sigma^2 = 1.0$. For the correctly specified exposure model, the average out-of-sample prediction accuracy is $\bar{R}^2_W = 0.74$ and the health effect estimation standard deviation is 0.12 with a bias of −0.007 (95% CI = −0.008 to −0.006). Corresponding statistics for the misspecified exposure model are $\bar{R}^{2'}_W = 0.49$ and health effect estimation standard deviation 0.21 with a bias of −0.001 (95% CI = −0.002 to 0.0006). The standard error of $\hat{\alpha}_3$ for the correctly specified model is 0.41, and $\hat{\alpha}_3$ is statistically significant in all simulations.

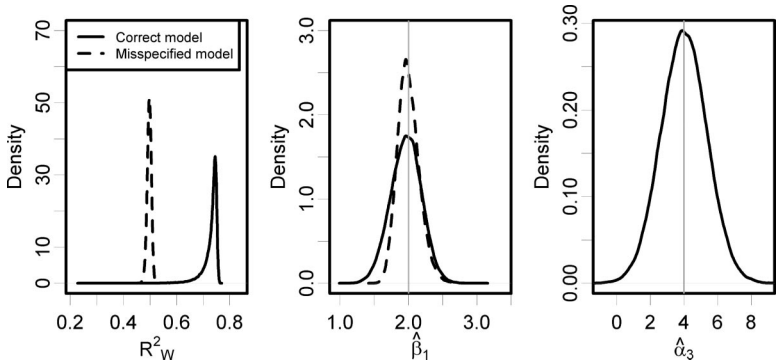

FIGURE 2. Results from 80,000 Monte Carlo simulations with $N = 10{,}000$, $N^* = 100$, and $\sigma^2 = 0.1$. For the correctly specified exposure model the average out-of-sample prediction accuracy is $\bar{R}^2_W = 0.73$ and the health effect estimation standard deviation is 0.23 with a bias of −0.035 (95% CI = −0.037 to −0.034). Corresponding statistics for the misspecified exposure model are $\bar{R}^{2'}_W = 0.50$ and health effect estimation standard deviation 0.16 with a bias of 0.001 (95% CI = −0.0003 to 0.002). The density plot for R^2_W shows some small outliers for the full model, but the prediction accuracy is better than for the misspecified model in all but 144 of the 80,000 simulations. The standard deviation of $\hat{\alpha}_3$ for the correctly specified model is 1.37, and $\hat{\alpha}_3$ is statistically significant in 83% of simulations.

because classical-like error induces bias while Berkson-like error does not.

It turns out that for $\sigma^2 = 0.1$ in the monitoring data, we get relatively variable estimates of α_3 when using the full exposure model, while still improving out-of-sample prediction accuracy at subjects' locations. This results in substantial classical-like measurement error that (for sufficiently large N) is more important than the additional Berkson error that is introduced by omitting the corresponding geographic covariate. There is very little bias in any of our simulations, and so the dominant classical-like error primarily results in more variable estimates of β_X.

IMPLICATIONS FOR FUTURE RESEARCH

We have shown a class of examples in which more accurate exposure prediction does not lead to improved health effect estimation. It bears emphasis that this does not result from overfitting the exposure model, at least not as overfitting is traditionally understood for prediction models.[13] In all cases, using the correctly specified model that includes all 3 geographic covariates leads to improved prediction accuracy, as measured by out-of-sample R^2_W evaluated at subjects' locations.

Our findings have important implications for the design and analysis of environmental epidemiologic studies. Development of models for exposure prediction and health effect estimation should be considered simultaneously, in contrast with the current practice of first selecting an exposure model to optimize prediction accuracy and then using the resulting predictions for health effect estimation. Recent papers that address measurement error in air pollution cohort studies represent progress in this direction.[11,12,14,17] Our results do

© 2011 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

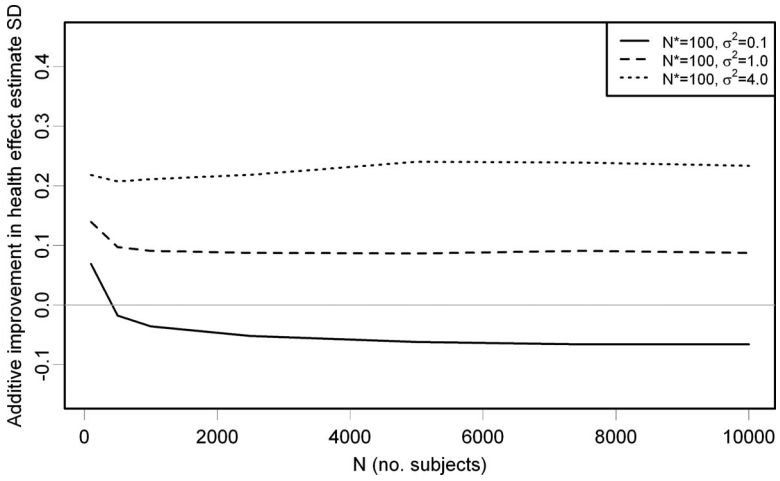

FIGURE 3. Results from 5000 Monte Carlo simulations with $N^* = 100$, $\sigma^2 = 0.1$, 1.0, 4.0, and N ranging from 100 to 10,000. The vertical axis shows the difference between standard deviation (SD) of $\hat{\beta}_X$ from the misspecified and correct exposure models. A positive difference indicates that the correctly specified model is more efficient. For all values of σ^2, the average exposure model prediction accuracies are $\bar{R}^2_W$ between 0.73 and 0.75 and $\bar{R}^{2'}_W$ between 0.49 and 0.50.

not necessarily suggest employing a joint statistical estimation model for the exposure and health parameters in which the health data would influence estimation of the exposure model parameters. The issue we have highlighted relates more directly to model selection than to parameter estimation.

There is extensive literature on penalization and other methods for optimizing accuracy of prediction models,[13] but these techniques are not directly applicable because better prediction accuracy may induce less precise health effect estimation. New statistical methodology is needed to select exposure models to optimize efficiency of health effects inference, perhaps involving alternative forms of penalization that account for the structure in both the monitoring and health outcome data. It is also worth exploring asymptotic methods to estimate the bias and variance of $\hat{\beta}_X$ to select optimal geographic covariates, particularly when there is a relatively large number of monitoring locations compared with the geographic covariates.

The relative benefits of various air-pollution exposure models depend on the variability of geographic covariates in the subject population and monitor locations, and on the size of the cohort. It is evident that study design can be improved by accounting for statistical issues at the intersection of exposure prediction and health effect estimation. All else being equal, it is preferable to design an exposure monitoring campaign to maximize the variability of pertinent geographic covariates across monitor locations. An asset allocation-algorithm may be useful for optimizing the monitoring design to predict exposures in an epidemiology study with known subjects' locations.[18]

We have considered only the relatively simple setting of a linear disease model with an exposure model that is land-use regression with independent geographic covariates. Even in this case, we have shown that more accurate exposure prediction does not necessarily lead to improved health effect estimation. We expect that similar phenomena can occur in other settings, but further research is needed to identify general conditions and assess the implications of more complex situations.

ACKNOWLEDGMENTS

We thank 3 anonymous referees for their valuable suggestions and Sverre Vedal for helpful comments on a draft of this manuscript.

REFERENCES

1. Yanosky JD, Paciorek CJ, Suh H. Predicting chronic fine and coarse particulate exposure using spatio-temporal models for the northeastern and midwestern United States. *Environ Health Perspect*. 2009;117:522–529.
2. Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*. 2010;21:606–631.
3. Brauer M. How much, how long, what, and where: Air pollution exposure assessment for epidemiologic studies of respiratory disease. *Proc Am Thoracic Soc*. 2010;7:111–115.
4. Fanshawe TR, Diggle PJ, Rushton S, et al. Modelling spatio-temporal variation in exposure to particulate matter: a two-stage approach. *Environmetrics*. 2008;19:549–566.
5. Su JG, Jerrett M, Beckerman B, Wilhelm M, Ghosh JK, Ritz B. Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy. *Environ Res*. 2009;109:657–670.
6. Jerrett M, Arain A, Kanaroglou P, et al. A review and evaluation of intraurban air pollution exposure models. *J Expo Anal Environ Epidemiol*. 2005;15:185–204.
7. Hoek G, Beelen R, de Hoogh K, et al. A review of land-use regression models to assess spatial variation in outdoor air pollution. *Atmos Environ*. 2008;42:7561–7578.
8. Jerrett M, Burnett RT, Ma R, et al. Spatial analysis of air pollution mortality in Los Angeles. *Epidemiology*. 2005;16:727–736.
9. Kunzli N, Jerrett M, Mack WJ, et al. Ambient air pollution and atherosclerosis in Los Angeles. *Environ Health Perspect*. 2005;113:201–206.
10. Puett RC, Hart JE, Yanosky JD, et al. Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. *Environ Health Perspect*. 2009;117:1697–1701.
11. Szpiro AA, Sheppard L, Lumley T. Efficient measurement error correction for spatially misaligned data. *Biostatistics*. In Press.
12. Gryparis A, Paciorek CJ, Zeka A, et al. Measurement error caused by

 © 2011 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

spatial misalignment in environmental epidemiology. *Biostatistics*. 2009;10:258–274.
13. Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning*. New York: Springer; 2001.
14. Kim SY, Sheppard L, Kim H. Health effects of long-term air pollution: influence of exposure prediction methods. *Epidemiology*. 2009;20:442–450.
15. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2010. ISBN 3-900051-07-0.
16. Carroll RJ, Ruppert D, Stefanski LA, et al. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC; 2006.
17. Madsen L, Ruppert D, Altman NS. Regression with spatially misaligned data. *Environmetrics*. 2008;19:453–467.
18. Kanaroglou PS, Jerrett M, Morrison J, et al. Establishing an air pollution monitoring network for intraurban population exposure assessment: A location-allocation approach. *Atmos Environ*. 2005; 39:2399–2409.

© 2011 Lippincott Williams & Wilkins

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.