American Journal of Epidemiology
Vol. 164, No. 1
DOI: 10.1093/aje/kwj150
Advance Access publication April 19, 2006
Copyright © 2006 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved; printed in U.S.A.

Original Contribution

Bayesian Modeling of Air Pollution Health Effects with Missing Exposure Data

John Molitor, Nuoo-Ting Molitor, Michael Jerrett, Rob McConnell, Jim Gauderman, Kiros Berhane, and Duncan Thomas

From the Department of Preventive Medicine, University of Southern California, Los Angeles, CA.

Received for publication June 8, 2005; accepted for publication January 13, 2006.

The authors propose a new statistical procedure that utilizes measurement error models to estimate missing exposure data in health effects assessment. The method detailed in this paper follows a Bayesian framework that allows estimation of various parameters of the model in the presence of missing covariates in an informative way. The authors apply this methodology to study the effect of household-level long-term air pollution exposures on lung function for subjects from the Southern California Children's Health Study pilot project, conducted in the year 2000. Specifically, they propose techniques to examine the long-term effects of nitrogen dioxide (NO_2) exposure on children's lung function for persons living in 11 southern California communities. The effect of nitrogen dioxide exposure on various measures of lung function was examined, but, similar to many air pollution studies, no completely accurate measure of household-level long-term nitrogen dioxide exposure was available. Rather, community-level nitrogen dioxide was measured continuously over many years, but household-level nitrogen dioxide exposure was measured only during two 2-week periods, one period in the summer and one period in the winter. From these incomplete measures, long-term nitrogen dioxide exposure and its effect on health must be inferred. Results show that the method improves estimates when compared with standard frequentist approaches.

air pollution; Bayesian analysis; bias (epidemiology)

Abbreviations: FEV_1, forced expiratory volume in 1 second; FVC, forced vital capacity.

Adequate assessment of environmental exposures that vary within communities in population-based epidemiologic studies is limited by the expense involved in obtaining measurements at multiple locations, often for prolonged periods. An example is air pollution, for which studies of chronic health effects have traditionally relied on continuous measurements made at central-site monitors (1–4). Although successful for demonstrating initial associations, central-site measurements fail to capture the large variability in exposures within communities that occurs near roadways and stationary sources (5). Recent studies have shown that within-community exposure gradients may be associated with larger health effects than the between-community exposures used in earlier studies (6–8).

This transition to studies using within-community exposure gradients raises measurement and statistical issues. In particular, local-scale monitoring information is needed to calibrate and confirm exposure assignments, and there is increased potential for measurement error in estimated exposure. With the large uncertainty in exposure estimates, questions remain about the validity of results from health effects studies that use exposure surrogates based on incomplete information, such as road buffers or models fitted with sparse monitoring data. The combination of large initial health effects and the heightened potential for errors has prompted researchers to identify the development of models for assessing air pollution exposure within cities as a priority for future research (9, 10).

Correspondence to Dr. John Molitor, Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90089-9011 (e-mail: jmolitor@usc.edu).

Downloaded from http://aje.oxfordjournals.org/ at University of North Texas on January 5, 2017

This paper focuses on developing improved exposure models for assigning within-community variation in ambient traffic pollution to the Children's Health Study, a prospective study of schoolchildren in 12 communities in southern California (1–4, 11–13). The Children's Health Study has already yielded important findings about the relation between lung function growth and air pollution exposure by between-community comparisons, but we now wish to address the question of whether intraurban gradients in exposure contribute to the relations observed with central-site comparisons (14). The methods proposed utilize well-studied measurement error techniques (15, 16) in estimating these unobserved exposures.

MATERIALS AND METHODS

We propose an alternative method to address these two measurement error problems in exposure assessment: 1) estimating "true" long-term exposure from flawed short-term measurements and 2) estimating missing measurement (or other covariate) information in locations where no measurements have been made. We will utilize Bayesian Markov chain Monte Carlo estimation methods (17) that model the process through which the unobserved exposures are estimated. Markov chain Monte Carlo methods can fit an entire multilevel model as a unit, properly taking into account parameter estimation uncertainty at each level of the model.

Data

We illustrate the methods using data from the Southern California Children's Health Study. In the Children's Health Study, continuous, long-term central-site measurements of air pollution were made in multiple communities. Two seasonal short-term household-level measurements were made at a subset of study participant residences within communities to characterize local deviations from the community-specific control site measurements at the same times. Using the proposed methods, we estimated household-level long-term residential exposures for persons with and without seasonal exposure measurements, and these estimates were used to evaluate the effect of air pollution on lung function in children. This outcome has been well studied by use of central monitors for ecologic, between-community comparisons (1, 4, 13, 18, 19), but there has been little study of the effect of residential exposures that vary within communities. The above-mentioned estimation procedures were performed in a unified Bayesian framework.

The 12 Children's Health Study communities were selected to have different levels of exposure to ozone (O_3), nitrogen dioxide (NO_2), and atmospheric particulate matter with a diameter of 10 μm or less (PM_{10}), based upon previously available air-monitoring data. (Note that the present analysis is limited to 11 communities, excluding the mountain community of Lake Arrowhead, for which home addresses could not be geocoded accurately.) Extensive questionnaire data on demographic characteristics, health outcomes, activities, and housing characteristics were obtained at study entry and updated annually. Lung function, height, and weight measurements were made at study entry and annually up to high school graduation. Lung function was measured by spirometry (13). For these analyses, we have focused on forced vital capacity (FVC) and forced expiratory volume in 1 second (FEV_1) as the outcomes of interest, as these have proven in our analyses to be sensitive indicators of lung response to regional air pollutants. These indicators of lung function evaluate the effects on both lung volume (FVC) and airway flow (FEV_1).

The locations of all schools and home addresses were geocoded. Estimates of average nitrogen dioxide exposure from traffic on freeways and major surface streets, as well as predicted pollution exposures from mobile surfaces, using the CALINE4 line source dispersion model (20) were obtained.

The household pollution data that we have analyzed come from a study conducted in the year 2000, in which outdoor measurements of nitrogen dioxide concentrations were made at 233 homes of Children's Health Study children during one 2-week period in the summer and one 2-week period in the winter. Nitrogen dioxide was measured because vehicular traffic is the major source of local variability in nitrogen dioxide in urban areas without major industrial sources, and there is some evidence that nitrogen dioxide is a good indicator of local variation in other traffic-related pollutants, including fine and ultrafine particulate matter (5, 10, 21). Unlike particulate pollution, nitrogen dioxide can be measured unobtrusively at homes at a cost that allows for multiple simultaneous samples.

Model

The aims of this paper are to model the determinants of local variation in outdoor concentrations of nitrogen dioxide in the Children's Health Study in relation to traffic patterns and to use this model to estimate long-term nitrogen dioxide exposure. These estimated determinants will be used as covariates in a model for lung function. In this context, nitrogen dioxide serves as a proxy for local traffic pollution exposure. Our approach to this problem is grounded in the statistical literature on exposure measurement error (15, 16, 22).

This analysis involves $C = 11$ towns, denoted $c = 1, \ldots, C$, with $i = 1, \ldots, N_c$ individuals per town. Measurements were made during two seasons, denoted $j = 1, 2$. The lung function measurements are denoted Y_{ci}, the observed household-level exposure measurements Z_{cij}, and the unobserved annual household-level concentrations X_{ci}. Here, we let the subscript * indicate the average over the corresponding index. We let $\boldsymbol{W}_{ci}$ denote household-level exposure variables that influence local concentrations, such as distance from the nearest freeway and predicted nitrogen dioxide concentration from the CALINE4 (20) line-source dispersion model. The CALINE4 model incorporates distances from traffic densities on all major nearby roads along with the frequency distribution of wind speeds and direction. Personal covariates that influence lung function directly, such as smoking, asthma, and respiratory illness at the time of lung function measurement, are denoted $\boldsymbol{V}_{ci}$. Our analytical framework consists of the following three

Downloaded from http://aje.oxfordjournals.org/ at University of North Texas on January 5, 2017

submodels, hereafter called the *disease*, *exposure*, and *measurement* models, respectively.

$$Y_{ci} = A_c + \alpha(X_{ci} - X_{c*}) + \boldsymbol{V}_{ci}\boldsymbol{\eta}' + e_{ci} \quad \text{(disease model)} \quad (1)$$

$$X_{ci} = B_c + \boldsymbol{W}_{ci}\boldsymbol{\delta}' + f_{ci} \quad \text{(exposure model)} \quad (2)$$

$$Z_{cij} = X_{ci} + \gamma(P_{cj} - P_{c*}) + g_{cij} \quad \text{(measurement model)} \quad (3)$$

We also model community-level random effects for the disease and exposure models as $A_c = \omega_0 + \omega_1 P_{c*} + h_c$ and $B_c = \beta_0 + \beta_1 P_{c*} + k_c$. Here, we let e_{ci}, f_{ci}, g_{cij}, h_c, and k_c represent normally distributed error terms with standard deviations σ_e, σ_f, σ_g, σ_h, and σ_k, respectively.

Equation 1 models the effect of household-level long-term nitrogen dioxide exposure along with the effects of various personal-level covariates on lung function. Equation 2 uses various household-level covariates to help predict the long-term nitrogen dioxide exposure. Since we only have household-level nitrogen dioxide measurements from two time periods throughout the year for the 233 homes in the study, we model the long-term level of nitrogen dioxide exposure for individual i in community c, X_{ci}, using equation 3. This modeled household-level long-term nitrogen dioxide exposure will simply be referred to as "modeled" nitrogen dioxide exposure in the rest of this paper, but it should be understood to estimate long-term exposure to nitrogen dioxide in the home.

The overall model is depicted in figure 1. The joint distribution of all parameters in the model can be expressed as follows:

$$f(\boldsymbol{Y}|A_c,\alpha,\boldsymbol{X},\boldsymbol{V},\boldsymbol{\eta},\sigma_e)f(\alpha)f(\boldsymbol{\eta})f(\sigma_e)f(A_c|\boldsymbol{\omega},\boldsymbol{P},\sigma_d)f(\boldsymbol{\omega})f(\sigma_d)$$
$$\times f(\boldsymbol{X}|B_c,\boldsymbol{W},\boldsymbol{\delta},\sigma_f)f(\boldsymbol{\delta})f(\sigma_f)f(B_c|\boldsymbol{\beta},\boldsymbol{P},\sigma_i)f(\boldsymbol{\beta})f(\sigma_i)$$
$$\times f(\boldsymbol{Z}|\boldsymbol{X},\gamma,\boldsymbol{P},\sigma_g)f(\gamma)f(\sigma_g).$$

Bayesian estimation procedures

We set our model in a Bayesian framework and estimate parameters using the Markov chain Monte Carlo method, Gibbs sampling (17). One advantage of these procedures is that missing data can be handled in a natural way. In this technique, each parameter in the model is sampled from its full conditional distribution, that is, the distribution obtained by conditioning on all the other unknowns in the model. Parameters, missing covariates, and latent variables are, in a Bayesian context, seen as random variables, each of which can be estimated using Bayesian parameter estimation techniques. For example, if we let boldface indicate vectors of household-level elements, such as $\boldsymbol{X} = (X_{11}, X_{12}, \ldots, X_{CI})$, then, in each cycle of the Gibbs sampler, values of X_{ci} are updated from

$$f(X_{ci}|\bullet) \propto f(Y_{ci}|X_{ci},\bullet)f(X_{ci}|\boldsymbol{W}_{ci},\boldsymbol{P}_{c*},\bullet)f(\boldsymbol{Z}_{ci}|X_{ci},\boldsymbol{P}_{ci},\bullet),$$

where $\bullet$ indicates all relevant parameters for the distribution in question. Even if observed household-level nitrogen dioxide measurements, Z_{cij}, are missing for a particular X_{ci}, estimates can be obtained by the current estimates of all the parameters in the model. In turn, each parameter in the model is updated by use of all available data and current updates of all other parameters in the model. In other words, the relation between various facets of the model such as $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{Z}$ can be obtained for the data that are available, and these relations can be used to impute values for covariates that are not available.

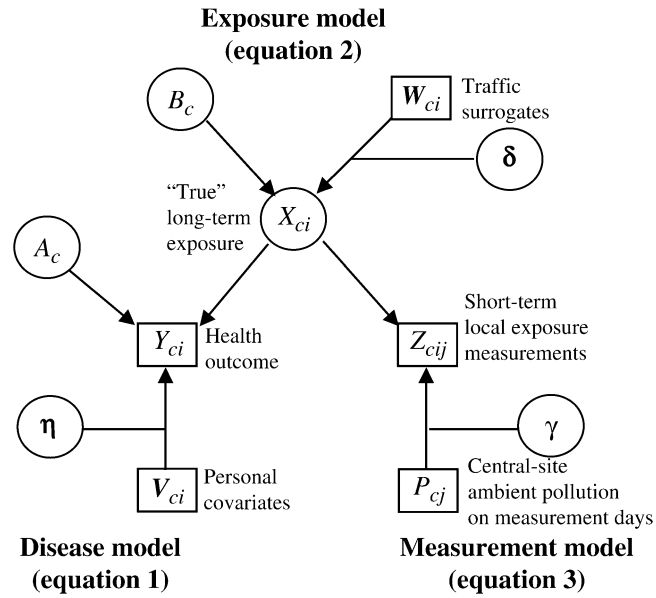

FIGURE 1. Schematic representation of the overall model used to analyze data in the Children's Health Study pilot project, conducted in the year 2000. Note that, while not represented in the graph, the random intercepts A_c and B_c are modeled by use of central-site exposure measurements P_{c*}.

All Markov chain Monte Carlo analyses were conducted using the WinBUGS software package (23). (This program is available upon request from the first author of this paper.) The Bayesian models were run for a burn-in of 20,000 iterations, followed by 100,000 iterations that were stored to compute posterior distributions. Diffuse priors were used on all parameters. Regression parameters were assigned $N(0, \tau_N)$ priors (here, τ_N denotes precision), with $\tau_N = 1.0e-12$. Variance components were given flat uniform priors, $U(0, \tau_U)$ as suggested by Gelman et al. (24), as opposed to conjugate priors. We used $\tau_U = 100$ to define our vague prior for all variance components. Throughout the analyses, all measures of nitrogen dioxide, both estimated and observed, distance to the nearest freeway, and CALINE4-predicted nitrogen dioxide, as well as the outcome, Y_{ci}, were measured on a log scale to satisfy the normality assumptions of the models.

Prior sensitivity was examined by making the priors tighter, that is, less vague. For instance, the effects of using larger values of τ_N (e.g., $\tau_N = 1.0e-6$) and smaller values of τ_U (e.g., $\tau_U = 10$) were examined. In general, the results were quite robust to changes in these parameter values as long as the prior specifications were sufficiently vague. The sampler exhibited good convergence properties as well.

Downloaded from http://aje.oxfordjournals.org/ at University of North Texas on January 5, 2017

Time-series plots of posterior parameter quantities indicated that the mixing of the sampler was extremely good. Multiple chains were run using different starting values, but the end results were nearly identical for all chains. These features indicate that the total of 120,000 iterations used was much more than necessary to achieve convergence.

The interpretation of parameter estimates obtained when using these kinds of log-log models corresponds to what is commonly known in the regression literature as *elasticity*. The coefficient in front of a particular covariate is interpreted as the percent change in the response, Y, corresponding to a 1 percent change in the value of the covariate, X, assuming that everything else in the model is held constant. For example, if the model is $\log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon$, then if $\beta_1 = 0.2$, a 10 percent increase in the value of X will lead to a 2 percent change of 0.02 in the response Y.

Regression model comparisons

The data were also fitted to frequentist regression models for comparison purposes. The basic model setup regresses the observed seasonal nitrogen dioxide measurements on traffic covariates and then uses this equation to predict modeled nitrogen dioxide exposure, $\hat{X}_{ci}$. These predictions are then used as covariates in the model for lung function, Y_{ci}, in the second stage. The model setup consists of a measurement and exposure models (equation 4) and a disease model (equation 5):

$$\bar{\boldsymbol{Z}}_{ci} = \beta_0 + \boldsymbol{W}_{ci}\boldsymbol{\delta}' + e_{1ci} \tag{4}$$

$$Y_{ci} = A_c + (\hat{X}_{ci} - \hat{X}_{c*})\boldsymbol{\alpha}' + \boldsymbol{V}_{ci}\boldsymbol{\eta}' + e_{2ci}. \tag{5}$$

Here again, $\boldsymbol{V}$ denotes personal covariates, $\boldsymbol{Z}$ denotes local nitrogen dioxide measurements, $A_c = \omega_0 + \omega_1 P_{c*} + h_c$, and residual errors follow normal distributions centered at zero.

The challenge in using the above model setup is to find a way to predict the modeled measures of nitrogen dioxide exposure, X_{ci} in model 5, from observed seasonal measures, Z_{ci} in model 4, in such a way that uncertainty in the estimation process is properly taken into account. We will compare our previously described Bayesian approach with three frequentist regression approaches. In all three approaches, we model nitrogen dioxide exposure in the second-stage model (equation 5) by using the fitted values obtained from the first-stage model (equation 4). The three approaches are described below.

Naive model. Rather than modeling modeled nitrogen dioxide exposure, we simply insert that seasonal outcome from equation 5 into equation 4 as

$$Y_{ci} = A_c + (\boldsymbol{Z}_{ci} - \bar{\boldsymbol{Z}}_c)\alpha' + \boldsymbol{V}_{ci}\boldsymbol{\eta}' + e_{ci}. \tag{6}$$

Weighted single-imputation regression. Nitrogen dioxide exposure was modeled by use of weighted regression techniques, that is, $\hat{X}_{ci}$ in equation 5 was set equal to $E(\bar{Z}_{ci}) = \hat{\beta}_0 + \boldsymbol{W}_{ci}\hat{\boldsymbol{\delta}}'$.

Multiple imputation. A total of K sets of multiple first-stage nitrogen dioxide predictions were generated for each person. These outcomes were sampled from the fitted distribution obtained in model 4, namely, $N(\hat{X}_{ci}, \hat{\sigma}^2)$, with $\hat{\sigma}^2$ denoting the estimated residual error. New fitted values were obtained for each set of generated outcomes, and these fitted values were imputed into model 5, with each set of fitted values imputed in the same manner as was done in the single-imputation case. The parameter estimates obtained from the corresponding K regressions were then combined by use of multiple-imputation methods (25, 26). Results obtained from multiple regressions based on imputed values are combined with the PROC MIANALYZE procedure in SAS software (27). For explanations of multiple-imputation methods, refer to the books by Rubin (25) and Schafer (26). The default value of $K = 5$ imputations was used.

Comparison of results from these models serves as a useful benchmark for our Bayesian methods. We chose to estimate these frequentist regression models using the SAS software package (27). This way, we can compare the results obtained from our unified Bayesian model with those obtained from a standard frequentist analysis using a standard statistical software package.

RESULTS

The results depicted in table 1 show that the effects of modeled nitrogen dioxide on lung function for FVC are statistically significant and that the effects of nitrogen dioxide on FEV_1 are marginally significant. Figure 2 displays the posterior distributions for the parameter α, the effect of modeled nitrogen dioxide exposure, X_{ci}, on lung function for both FVC and FEV_1, and demonstrates that modeled nitrogen dioxide exposure clearly affects lung function in the negative direction. For FVC, the probability $\Pr(\alpha > 0)$ equals $p = 0.019$. Similarly, for FEV_1, $p = 0.032$. (These quantities can be thought of as Bayesian one-sided p values.) The distance to the nearest freeway provides little information in estimating the level of modeled nitrogen dioxide exposure when CALINE4-predicted nitrogen dioxide is included in model 2, since distance to freeway is one of the factors included in the calculation of CALINE4-predicted nitrogen dioxide. However, the coefficients take the expected sign, as an increase in distance from a freeway is associated with a decrease in predicted modeled nitrogen dioxide exposure.

Results obtained by use of frequentist models 4 and 5 are summarized, along with the Bayesian results, in table 1. The results pertaining to the effect of modeled nitrogen dioxide exposure on lung function obtained from the Bayesian model are the only ones that show significance when FVC is used as an outcome. Moreover, Bayesian credible intervals pertaining to this modeled nitrogen dioxide effect, the effect of primary interest, are narrower than all the confidence intervals obtained using frequentist approaches. It is also important to note that results obtained from the frequentist models relating to the effects of traffic-related covariates on seasonal nitrogen dioxide exposures were calculated without regard to the measure of lung function used in the second-stage model 5. Consequently, the results here are the same regardless of the measure of lung function used. The Bayesian model uses a unified approach in estimating model

Downloaded from http://aje.oxfordjournals.org/ at University of North Texas on January 5, 2017

TABLE 1. Results for Bayesian and frequentist models, Children's Health Study pilot project, conducted in the year 2000*

	FEV_1† (% change)		FVC† (% change)	
	Posterior mean	95% confidence interval	Posterior mean	95% confidence interval
Long-term nitrogen dioxide‡,§				
Bayesian model	−0.145	−0.308, 0.009	−0.138	−0.275, −0.006
Frequentist models				
Naïve approach	−0.048	−0.129, 0.032	−0.058	−0.129, 0.013
Weighted regression	−0.210	−0.440, 0.020	−0.145	−0.341, 0.051
Multiple imputation	−0.148	−0.311, 0.016	−0.127	−0.269, 0.014
	True nitrogen dioxide (% change)		True nitrogen dioxide (% change)	
	Posterior mean	95% confidence interval	Posterior mean	95% confidence interval
Distance to freeway‡,§,¶				
Bayesian model	−0.027	−0.066, 0.014	−0.027	−0.068, 0.015
Weighted regression/multiple imputation#	−0.040	−0.081, 0.001	−0.040	0.081, 0.001
Total CALINE4**-predicted nitrogen dioxide‡,§				
Bayesian model	0.239	0.163, 0.320	0.239	0.161, 0.317
Weighted regression/multiple imputation#	0.221	0.139, 0.303	0.221	0.139, 0.303

* Long-term nitrogen dioxide effects were computed for the Bayesian model and for frequentist models by use of the naive approach, weighted regression, and multiple imputation.

† FEV_1, forced expiratory volume in 1 second; FVC, forced vital capacity.

‡ Adjusted for age centered at 10 years, height, body mass index, gender, cohort, race, Hispanic ethnicity, asthma status, any tobacco smoking in the last year, exercise, and any respiratory illness on the day of testing.

§ Centered at the town-specific mean.

¶ Distance (m)/1,000.

The effects here are the same for FVC and FEV_1, since the naive model does not include measures of lung function in the second-stage model to estimate first-stage effects on nitrogen dioxide.

** CALINE4, a dispersion model for predicting air pollution concentration (20).

parameter values, so the measure of lung function used affects parameter estimates at all modeling stages.

A comparison of estimated modeled nitrogen dioxide exposures obtained from the multilevel model with raw seasonal and central-site estimates is depicted in table 2 with FVC used as the outcome. Results for FEV_1 are similar. With the Bayesian model, each modeled nitrogen dioxide variable, X_{ci}, has its own unique posterior distribution. Individual-level means and standard deviations obtained from these posterior distributions were obtained and then used to provide the modeled nitrogen dioxide estimates for all persons in a particular community. In other words, the average of all posterior means and the average of all posterior variances for the persons in a town were computed. Raw estimates were obtained simply by averaging seasonal measurements. At the community level, the nitrogen dioxide estimates obtained from the Bayesian model correspond roughly to the average of the seasonal and central-site measurements.

It is informative to relate the range of intercommunity estimates of modeled nitrogen dioxide listed in table 2 to the parameter estimates obtained in table 1. For example, in the town of Riverside, California, the estimated modeled nitrogen dioxide exposure across persons was 36.31 parts per billion (ppb) with a standard deviation of 5.510 ppb. An interval of ±2 standard deviations from the mean represents the range of nitrogen dioxide exposure present in this community, excluding outliers. The lower endpoint of such an interval is $L = 36.31 - 2 \times 5.510 = 25.29$ ppb, and the upper endpoint is $U = 47.33$ ppb. An increase in modeled nitrogen dioxide exposure from L to U represents an increase of $(47.33 - 25.29)/25.29 = 87$ percent. From table 1, we see that the coefficient for modeled nitrogen dioxide is $a \approx -0.14$ when FVC is used as the outcome. This means that, if one were to compare a person having a modeled nitrogen dioxide exposure equal to $L = 25.29$ ppb with a person having exposure equal to $U = 47.33$ ppb, one would expect to see, on average, a decrease in lung function of $87 \times 0.14 = 12$ percent, assuming that all the other parameters of the model remain fixed.

DISCUSSION

The proposed modeling approaches use measurement error techniques in a Bayesian setting to estimate modeled residential nitrogen dioxide exposure at homes where

Downloaded from http://aje.oxfordjournals.org/ at University of North Texas on January 5, 2017

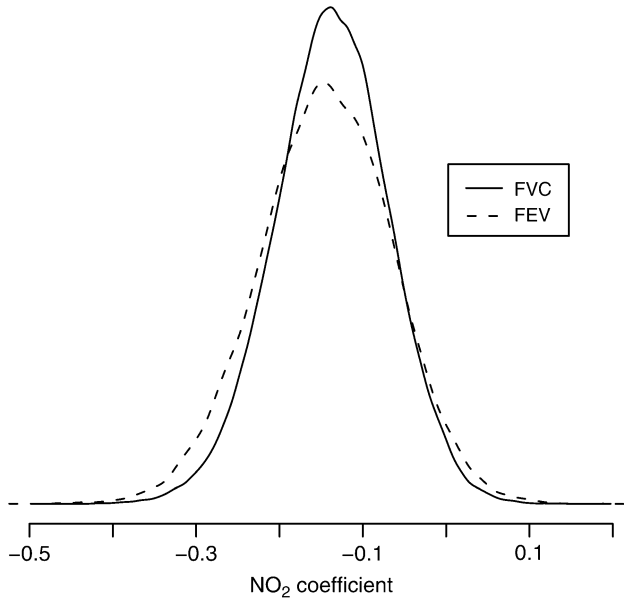

FIGURE 2. Posterior distributions for the effect of true nitrogen dioxide (NO_2) on lung function for the Children's Health Study pilot project, conducted in the year 2000. FVC, forced vital capacity; FEV, forced expiratory volume.

exposure was measured with error. The estimates relied on residential measurements that contained error because of the short duration of sampling, on predictors of exposure from widely used dispersion models, and on modeled community central-site air pollution measurements. Similar flawed indicators of local exposure are common in air pollution epidemiology and in studies of the effects of other exposures that are expensive to measure at many locations within communities. Therefore, this method provides an alternative modeling strategy that may have applicability to exposure assessment in many settings.

The approach is different because most previous studies have either used one exposure estimate to represent modeled exposure or simply weighted for the statistical uncertainty in the estimates (6, 7), rather than building the uncertainty into an integrated model. Alternative ad hoc approaches to imputing missing environmental exposures, such as using averages of covariate values, lack the integrated framework of the methods we propose. By use of an explicit model for missing values, a parameter estimation procedure such as Markov chain Monte Carlo can be utilized that explicitly takes these missing data into account. This integrated approach allows investigators to examine the specific effect of interest, namely, the effect of modeled nitrogen dioxide exposure on lung function.

The observed association of FEV_1 and FVC with the estimated residential exposure for all study participants suggests that variability in exposure to traffic-related pollutants may impair respiratory health. Few previous studies have examined the association of childhood lung function with variation within communities. A German study found significant deficits in peak expiratory flow rates and maximal expiratory flow rates associated with traffic counts near schools (28). Heavy traffic, especially truck traffic, within 300 m of homes in the Netherlands was associated with childhood deficits in both FVC and FEV_1 (29), but a more recent study by the same group, also examining the effects of heavy truck traffic on nearby roads, found no associations of traffic-related pollutants with lung function (30). In a panel study of the acute effects on exhaled nitrogen oxide, Koenig et al. (31) reported slightly larger associations for personal measurements of atmospheric particulate matter with a diameter of 2.5 μm or less ($PM_{2.5}$) than for central-site

TABLE 2. Comparison of raw seasonal and central-site estimates with estimated long-term nitrogen dioxide, when forced vital capacity is the outcome, Children's Health Study pilot project, conducted in the year 2000

California city	Nitrogen dioxide estimates (ppb)						
	Summer (Z_1)		Winter (Z_2)		Central site (P_{c*}) (mean)	Long term (X_{ci})	
	Mean	Standard deviation	Mean	Standard deviation		Mean	Standard deviation
Apline	19.960	4.996	19.027	7.673	12.742	16.399	1.440
Atascaero	12.070	4.662	15.708	8.230	12.541	13.298	1.190
Lake Elsinore	17.510	2.663	27.491	3.113	17.796	21.166	1.781
Lancaster	18.311	3.247	23.299	2.929	16.679	18.826	1.645
Lompoc	5.344	1.456	13.861	4.188	3.758	11.366	0.984
Long Beach	34.034	5.911	48.466	11.314	33.694	39.916	3.445
Mira Loma	37.174	4.335	48.471	3.735	24.532	37.884	3.171
Riverside	38.436	7.195	42.837	7.556	25.387	36.315	3.023
San Dimas	51.399	5.647	50.781	7.127	34.715	50.691	4.120
Santa Maria	13.215	4.600	17.479	3.247	11.484	15.184	1.326
Upland	46.317	7.172	35.708	5.881	38.271	39.124	3.370

Downloaded from http://aje.oxfordjournals.org/ at University of North Texas on January 5, 2017

monitor estimates. This study lends support to the idea that more precise measurements of exposure may uncover larger respiratory effects, but the panel design and personal exposure measurements cannot practically be applied to larger populations.

Our results are consistent with an emerging consensus from between-community comparisons that regional pollutants, such as fine particulate matter, which are measured at community monitors and which may more uniformly cover entire communities, cause deficits in lung function (1, 4, 18, 19). In addition, variation in exposure within communities has been associated with childhood asthma in some (32–36), but not all (28, 37–39), studies. Inconsistencies between studies may be explained in part by poor exposure assessment (5). Therefore, this approach may be useful for assessing the impact of local traffic on other Children's Health Study outcomes, such as asthma and lung function growth, and in other studies of the health effects of ambient air pollution.

This project is conceptually similar to numerous other studies that use the main-study/validation-substudy paradigm (40). Here, we provide a unified framework for the analysis of such studies by modeling simultaneously the relation between X and Z in the substudy data and by using this information to model the relation between Y and Z in the main study by integrating over the unobserved Xs. In our context, this would entail using the information from the pilot study analyzed in this paper to make inference regarding the effect of modeled nitrogen dioxide exposure on lung function for all persons in the entire Children's Health Study, based on the surrogate variable traffic density (and ultimately spatial correlations, as discussed below). Further investigation is required to examine the statistical power available to utilize the model derived from the substudy to impute exposures to subjects in the main study. This will be achieved through additional field measurements of nitric oxide, nitrogen dioxide, and ozone at over 1,000 homes. These spatially exhaustive measurements will be taken over four seasons. Once available, the new measurements will be utilized to extend the model into novel areas, including spatial autocorrelation parameters for assessing exposure and residual confounding, the assessment of multipollutant effects, and formal simulation of power requirements for extension from substudies to the entire cohort.

Beyond this application, the methods developed here may have widespread applicability for environmental health studies where incomplete, but potentially useful, exposure measurements are available. As epidemiologists move toward within-community studies of exposure, methods to utilize data sets with missing exposure data will likely become more important.

ACKNOWLEDGMENTS

This study was supported by the Southern California Particle Center and Supersite, the Southern California Environmental Health Sciences Center (grant 5P30 ES07048), and the National Institute of Environmental Health Sciences (grant 1 P01 ES011627). The authors acknowledge funding from the US Environmental Protection Agency Science to Achieve Results (STAR) Program (grant RD-83184501-0).

Conflict of interest: none declared.

REFERENCES

1. Gauderman WJ, Avol E, Gilliland F, et al. The effect of air pollution on lung development from 10 to 18 years of age. N Engl J Med 2004;351:1057–67.
2. Dockery DW, Pope CA, Xu X, et al. An association between air pollution and mortality in six U.S. cities. N Engl J Med 1993;329:1753–9.
3. McConnell R, Berhane K, Gilliland F, et al. Air pollution and bronchitic symptoms in Southern California children with asthma. Environ Health Perspect 1999;107:757–60.
4. Raizenne M, Neas LM, Damokosh AI, et al. Health effects of acid aerosols on North American children: pulmonary function. Environ Health Perspect 1996;104:506–14.
5. Jerrett M, Arain A, Kanaroglou P, et al. A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol 2004;15:185–204.
6. Hoek G, Brunekreef B, Goldbohm S, et al. Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. Lancet 2002;360:1203–9.
7. Finkelstein MM, Jerrett M, Sears MR. Traffic air pollution and mortality rate advancement periods. Am J Epidemiol 2004; 160:173–7.
8. Nafstad P, Haheim LL, Wisloff T, et al. Urban air pollution and mortality in a cohort of Norwegian men. Environ Health Perspect 2004;112:610–15.
9. Brunekreef B, Holgate ST. Air pollution and health. Lancet 2002;360:1233–42.
10. Brauer M, Hoek G, Vliet P, et al. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. Epidemiology 2003;14:228–39.
11. Berhane K, Gauderman WJ, Stram DO, et al. Statistical issues in studies of the long-term effects of air pollution: the Southern California Children's Health Study. Stat Sci 2004;19:414–19.
12. Peters JM, Avol E, Navidi W, et al. A study of twelve southern California communities with differing levels and types of air pollution. I. Prevalence of respiratory morbidity. Am J Respir Crit Care Med 1999;159:760–7.
13. Peters JM, Avol E, Gauderman WJ, et al. A study of twelve southern California communities with differing levels and types of air pollution. II. Effects on pulmonary function. Am J Respir Crit Care Med 1999;159:768–75.
14. Gauderman WJ, Avol E, Lurmann F, et al. Childhood asthma and exposure to traffic and nitrogen dioxide. Epidemiology 2005;16:737–43.
15. Carroll RJ, Ruppert D, Stefanski LA. Measurement error in nonlinear models. London, United Kingdom: Chapman & Hall, 1995.
16. Thomas DC, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. Annu Rev Public Health 1993;14:69–93.
17. Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain Monte Carlo in practice. London, United Kingdom: Chapman and Hall, 1996.
18. Dockery DW, Cunningham J, Damokosh AI, et al. Health effects of acid aerosols on North American children: respiratory symptoms. Environ Health Perspect 1996;104:500–5.

Downloaded from http://aje.oxfordjournals.org/ at University of North Texas on January 5, 2017

19. Gauderman WJ, Gilliland GF, Vora H, et al. Association between air pollution and lung function growth in southern California children: results from a second cohort. Am J Respir Crit Care Med 2002;166:76–84.
20. Benson P. CALINE4—a dispersion model for predicting air pollution concentration near roadways. Sacramento, CA: Office of Transportation Laboratory, California Department of Transportation, 1989.
21. Roorda-Knape MC, Janssen NA, Hartog J, et al. Traffic related air pollution in city districts near motorways. Sci Total Environ 1999;253:339–41.
22. Fuller W. Measurement error models. New York, NY: John Wiley & Sons, 1987.
23. Spiegelhalter D, Thomas A, Best N. WinBUGS version 1.4 user manual. Cambridge, MA: MRC Biostatistics Unit, 2003.
24. Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. Boca Raton, FL: Chapman & Hall/CRC, 1995.
25. Rubin DB. Multiple imputation for nonresponse in surveys. New York, NY: John Wiley & Sons, 1987.
26. Schafer JL. Analysis of incomplete multivariate data. London, United Kingdom: Chapman & Hall, 1997.
27. SAS language reference: dictionary, version 8. Cary, NC: SAS Institute, Inc, 2004.
28. Wjst M, Reitmeir P, Dold S, et al. Road traffic and adverse effects on respiratory health in children. BMJ 1993;307: 596–600.
29. Brunekreef B, Janssen NA, Hartog J, et al. Air pollution from truck traffic and lung function in children living near motorways. Epidemiology 1997;8:298–303.
30. Janssen NA, Brunekreef B, Vliet P, et al. The relationship between air pollution from heavy traffic and allergic sensitization, bronchial hyperresponsiveness, and respiratory symptoms in Dutch schoolchildren. Environ Health Perspect 2003; 111:1512–18.
31. Koenig JQ, Jansen K, Mar TF, et al. Measurement of offline exhaled nitric oxide in a study of community exposure to air pollution. Environ Health Perspect 2003;111:1625–9.
32. Venn A, Lewis S, Cooper M, et al. Local road traffic activity and the prevalence, severity, and persistence of wheeze in school children: combined cross sectional and longitudinal study. Occup Environ Med 2000;57:152–8.
33. Edwards J, Walters S, Griffiths RK. Hospital admissions for asthma in preschool children: relationship to major roads in Birmingham. Arch Environ Health 1994;49:223–7.
34. Hirsch T, Weiland SK, Mutius E, et al. Inner city air pollution and respiratory health and atopy in children. Eur Respir J 1999;19:669–77.
35. van Vliet P, Knape M, de Hartog J, et al. Motor vehicle exhaust and chronic respiratory symptoms in children living near freeways. Environ Res 1997;72:122–32.
36. Nicolai T, Carr D, Weiland SK, et al. Urban traffic and pollutant exposure related to respiratory outcomes and atopy in a large sample of children. Eur Respir J 2003;21:956–63.
37. English P, Neutra R, Scalfa R, et al. Examining associations between childhood asthma and traffic flow using a geographic information system. Environ Health Perspect 1999;107:761–7.
38. Waldron G, Pottle B, Dod J. Asthma and the motorways—one district's experience. J Public Health Med 1995;17:85–9.
39. Kramer U, Koch T, Ranft U, et al. Traffic-related air pollution is associated with atopy in children living in urban areas. Epidemiology 2000;11:64–70.
40. Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. Biometrics 1991;47:851–69.

Downloaded from http://aje.oxfordjournals.org/ at University of North Texas on January 5, 2017