Data Analysis using NHIS-EPA Linked Files: Issues with using Incomplete Linkage

Rong Wei[1], Van Parsons, Jennifer Parker and Yulei He

National Center for Health Statistics, Hyattsville, MD 20782

Abstract

The National Health Interview Survey (NHIS) collects individual health outcome data that represents the civilian non-institutionalized population of the United States. As the NHIS is based on a complex survey design which includes survey weighting and clustering factors, it is recommended that analyses should be implemented using methods which make use of available design variables. A special situation arises for analyses using NHIS-EPA (Environmental Protection Agency) linked data, a database available at the National Center for Health Statistics (NCHS) that geographically links NHIS data covering years 1985 to 2005 to EPA pollutant data. The available linkage only partially covers the geography sampled by the NHIS, and the "representativeness" of the linked NHIS-EPA component to the nation may be questionable. The present study focuses on issues related to the analysis of the linked NHIS-EPA data in a model-based framework that preserves many of the NHIS survey design features, but avoids some of the design-based structural complexities resulting from large amounts of missing data. First, comparisons of the linked and un-linked components of the NHIS are made to establish a possible "non-representativeness" of the linked component. While perhaps not representative of the nation as a whole, the EPA-linked data can still be viewed as a valuable source of information regarding associations between health and air pollutants. Random effects and Bayesian models which account for the survey design are presented as means to study associations between NHIS health and EPA pollutant variable when restricted to the partial NHIS-EPA data.

Key Words: air pollutant, health status, linked complex survey data, mixed effects model, model-based analysis

[1] *The findings and conclusions in this study are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention*.

1. Introduction

To study associations between health outcomes and air quality, data from the National Health Interview Survey (NHIS) covering years 1985 – 2005 were geographically linked to air quality data from the Environmental Prevention Agency (EPA). This linkage is discussed in Parker et al. (2008a), and this resource database is only available in the NCHS Research Data Center. These two independent designs for data collection have different structures in terms of national geographic coverage and in the nature of sampling. The NHIS is a probability sample considered representative of the population and of the geographical coverage of the population, while the available EPA collected data cannot be considered as geographically complete nor population representative in nature. Linkage-eligibility due to different data collection purposes leaves unlinked data in both data sources.

Design-based analyses for NHIS-EPA associations have been previously studied, e.g., design-based logistic regression (Parker et al. 2009) and sample weighting adjustment methods (Judson, Parker and Larsen, 2013). An assumption that the missing linkage-eligibility cases were at random was embedded in those studies. On the other hand, due to limited geographical coverage, these association analyses can be complicated by a potential "non-missing at random" structure of the EPA linked data. In addition, from the epidemiological perspective, estimation of associations between health measures and air quality variables using simple regression models might be further improved by accounting for NHIS design features in Bayesian analyses.

One of the purposes of this work is to provide data users some additional operating characteristics of this data resource and suggest some model-based procedures, e.g., random effects and Bayesian models. First, we consider the "representativeness" of the linked NHIS-EPA population component. As a specific case, we consider whether a general health measure is "equal" between the linked EPA part and corresponding unlinked part. Second, even if the linked and unlinked components have different characteristics, epidemiological focused analyses on the linked component are still valuable. We discuss how survey design features, i.e., sampling weights and sampling clusters, can be incorporated into a model-based association analyses. As our goal is to provide guidance to data users, we suggest both random effects and Bayesian models easily implementable in popular software packages.

2. Data and linkage eligibility

For the NHIS data, the health status given by NHIS respondents are dichotomized as "healthy" ("excellent", "very good" and "good") and "unhealthy" ("fair" and "poor"), and this variable is used as the health outcome in this study. Covariate variables for health status are four race/ethnic groups (Hispanic, non-Hispanic white, non-Hispanic black and others), two genders (male and female) and age (centered at 45 years old). The health status, covariates, and design features are complete with no missing values (a relatively few units with unknown health status variables are discarded). We refer to this essentially full NHIS data, before linkage, as the "complete data".

The EPA data includes annual average measures of 6 EPA air pollutants which are particulate matters (fine (PM2.5) and large (PM10)), carbon monoxide (CO), sulfur

dioxide (SO2), ozone (O3), and nitrogen dioxide (NO2). These are linked to NHIS respondents if monitors are present within 20 miles of the respondent's housing unit Census block group. These measures are used as air pollutant exposure variables. Only partial EPA data for each of these six air pollutants are linked to NHIS data. We refer to NHIS-EPA linked data as the "partial data". The coverage of these two data systems varies by year, but as an example, the year 2005 data linkage eligibility is shown in Table 1. As can be seen from this table, the availability of pollutant data varies by type of pollutant and coverage by county. For example, only a limited number of NHIS counties had lead measures, thus limiting the number of NHIS sample persons with complete data. As the variable "county" is a cluster, that represents a sampled design structure, the totality of which represent the nation, these units are needed to explain the randomness of the data. Any non-random missing counties present inferential problems resulting from data analyses.

Table 1. Data linkage eligibility in NHIS-EPA 2005

	Percentage of 2005 NHIS counties and individuals linked to EPA air data			
	County(Total NHIS 844)[1]		**Individual(Total NHIS 98,649)**	
Pollutant	**N**	**%**	**N**	**%**
Ozone	**627**	**74**	**79,767**	**80**
$PM_{2.5}$	**626**	**74**	**80,109**	**81**
PM_{10}	**462**	**55**	**66,776**	**68**
CO	**393**	**47**	**62,283**	**63**
NO_2	**351**	**42**	**57,472**	**58**
Lead	**113**	**13**	**21,243**	**22**

[1/]County contains some sample linked to EPA data

3. Population differences between complete and partial data

3.1 Health status on the partial data compared to the complete data

The health status outcome "unhealthy" is compared between the "partial data" and "compete data" by treating the partial data as a domain within the complete data and using logistic regression techniques to test for a domain effect. A generalized mixed effects model (using SAS PROC GLIMMIX) is applied separately to the complete data for each of the 6 air linkage-eligibility indicators respectively:

Logit(unhealthy) = Race + Sex + Age + (*single-Air –linkage- indicator)* + County (3.1.1)

This model was also applied to data with linkage indicators for all 6 air linkage-data presence:

Logit(unhealthy) = Race + Sex + Age + (*Simultaneous 6 air indicators)* + County (3.1.2)
The air linkage indicators are defined equal to 1 if health status and air data are linked, and equal to 0 if health status and air data are not linked; the variable "County" is treated as being a normally distributed random effect. For these models the design structure is partially covered by treating county as a one-stage clustering effect corresponding to PSU sampling and the covariates correspond to oversampling by race factors and poststratification variables.

Examples of results for Ozone and PM10 indicators using models (3.1.1) and (3.1.2) are given in Figure 1. The models were run separately for each of the linked years in the NHIS linked data and are represented on the x-axis. In general, we observed that the linkage parameter was highly significant under model (3.1.1). When using multiple indicators in model (3.1.2) there may be a high degree of collinearity among the 6 pollutant indicators, the magnitude of which may vary over the time intervals. Thus, the significance levels for a specific variable may be diminished. It should be noted that the pollutant effects could be confounded with other geographical variables linked to the available EPA data, e.g., urban/rural status. The significance of the link-indicator is indicative of the differences of the health status on the populations represented by the partial data and represented by a full NHIS.

3.2 Race, Gender and Age effects on Health Status using the complete data versus using only the linked EPA part data.

The models of equation (3.1.1) without the indicator parameter were run separately on the complete data and on the partial data specified by each of the 6 air pollutants. Gender effect estimates on the Ozone and Lead partial data on response health status are shown in Figure 2a and 2b, respectively. The gender effects appear consistently larger for the complete data. As the partial data is embedded within the complete data, a test of significance becomes difficult to perform, but there is an indication of a difference. Race and age effects were also studied and showed similar patterns with the effects for the complete data set dominating the effects for the corresponding partial data set.

While our analyses are limited in scope, we feel that these analyses suggest that health measures and some demographic characteristics may differ by a non-negligible amount on the linked and unlinked populations. Finer levels of analyses using additional covariates may suggest types of analyses that yield comparable results on linked and unlinked data. Our message is one must be cautious when making general inferences based on the linked data.

4. Association analysis

To study the possible associations between health outcomes and air pollutant measures, only the partial data may be directly used. Any design-based analysis considering all NHIS design factors are complicated by the patterns of missing linkage in the EPA linked data, and the design's structural integrity may be compromised. Thus, model-based alternatives that use survey features of weighting and clustering are suggested for

analyses. The NHIS uses multiple sampling procedures, but to simplify a model-based approach, we consider a conceptual two-stage sampling structure: county cluster sampling is the first random process followed by sampling individuals within sampled counties. Using a very basic generalized mixed effects modeling approach, the county-level clusters are treated as a random effect, and the individual response is treated as a binary variable. Design components of the NHIS may include differential sampling by race/ethnicity and weighting adjustments by race, gender and age distributions, so these three variables are included in the model as fixed covariates.

Models may be constructed as weighted and unweighted at the individual level. The weighted model uses an individual weight (denoted *wtsca* in the model expressions below) which is the NHIS survey weight, but scaled to a total effective sample size on the linked data. The individual weight is defined as

$wtsca_{,i} \equiv n_{total} \, (w_i / \sum_j w_j \,) \, / \, (CV^2(\underline{w}) + 1)$, where $\underline{w}$ are weights for the individuals in the partial data, and n_{total} is the unweighted total on the partial data.

These techniques are discussed in Section 4.4 of Korn and Graubard (1999) and were demonstrated for use on NHIS data in Wei and Parsons (2009) to help account for the complex design when using model-based approaches.

Four models were applied and select SAS®GLIMMIX outputs are displayed in Figure 3. The four models and corresponding Figure 3 labels (weighting status, symbol, color) are:

1) Covariate model: (nw, +, green)

$$\text{Logit(unhealthy)} = \beta_1\text{Race} + \beta_2\text{Sex} + \beta_3\text{Age} + \alpha\cdot(\text{single Air pollutant}) + \text{County}$$

2) Sample-adjusted weighted model: (w, ◊, blue)

$$\text{Logit}(\text{unhealthy}_{\text{wtsca}}) = \alpha\cdot(\text{single Air Pollutant}) + \text{County}$$

3) Covariate + Sample-adjusted weighted model: (w, o, red)

$$\text{Logit}(\text{unhealthy}_{\text{wtsca}}) = \beta_1\text{Race} + \beta_2\text{Sex} + \beta_3\text{Age} + \alpha\cdot(\text{single Air Pollutant}) + \text{County}$$

4) Simple model: (nw, x, orange)

$$\text{Logit(unhealthy)} = \alpha\cdot(\text{single Air Pollutant}) + \text{County}$$

where County ~ normal(0, σ^2_{cnty}) for all models.

Note, Model 3) makes the most use of the survey design features of clustering, differential weighting by race/ethnicity and poststratification.

Examples of results for the association study are given in Figure 3 using ozone- and PM10-linked data. The *T*-values for ozone and PM10 effects associated with "unhealthy" status are shown in Figures 3a and 3c, respectively; and the variance estimate for σ^2_{cnty} of the random effects variable, county, are shown in Figures 3b and 3d, respectively.

As can be seen, the four models track each other fairly well. The relative orders of magnitude of the estimates for σ^2_{cnty} over the four models tend to create an ordering of models: (3) < (1) < (2) < (4), (red < green < blue < orange), which is plausible as the modeled covariates should account for much of the variability. For the *T*-statistics the covariate models (1) and (3) track each other well as do the non-covariate models (2) and (4) track each other. However, the *T*-statistics do not display consistent significant levels over the time interval. Such behavior reinforces the notion that the linked NHIS-EPA data cannot be thought of as a "representative" design for some standard geographically time-linked population. As this paper is focusing on several simple models that may be first attempted by a typical data user, a careful analysis for a specific problem should involve careful covariate selection using perhaps sequential procedures and diagnostics to assess individual model fits. It should also be noted that the within-year tests are valid, but the linked geography may change in any time period. For example, in the year 1996 the NHIS was reduced by about 38%, and the annual EPA-linked counties may also change. With changes of counties, it is quite possible that other geographical related variables are confounding interpretations of the regressions. The nature of the NHIS-EPA linkage by year needs to be considered when making inference, and data users are encouraged to carefully explore possible confounding in any given NHIS-EPA data set.

5. Bayesian methods

Data users are increasingly using Bayesian methods for analyses. For the partial data we performed an analysis on the year 2005 data linkage using a Bayesian adaptation of model (1) in section 4.

Health Status ~ Bernoulli (p), where p = probability (unhealthy)
Logit (p) ~ normal ($BX + \alpha(Air), \Sigma$) ,
where B = vector of coefficients for race, sex and age, Σ is random effect of county;

Priors: Empirical Bayes parameters B and Σ which are estimated from complete data set;
α is non-informative, the variables B, α ~ normal and Σ ~ inverse gamma.

The SAS/MCMC procedure was applied to 10,000 out of 20,000 samples. Examples for select pollutant α variable posterior distributions are given in Figure 4a (ozone) and Figure 4b (PM10). An inspection of the quantiles for the posterior densities suggests that the PM10 pollutant is a significant factor on health status while ozone is not, i.e., at least 97.5% of the PM10 posterior probability is greater than 0.00169 while roughly 75% of the ozone posterior distribution is negative and 25% positive. Table 2 presents some comparisons between generalized mixed effects modeling using SAS GLIMMIX and SAS MCMC procedures. In general the Bayesian- and the random effects-approaches to modeling appear to be fairly consistent with respect to inference. The PM10 effects and the other covariate effects all appear to be of the same order of magnitude in the comparison. The ozone covariate comparisons looked comparable in magnitude except for the large difference for the ozone pollutant effect, -14.1 vs -0.8. This may be due to the Bayesian chain not having achieved a stationary state. As this Bayesian study is somewhat exploratory, additional work is necessary.

6. Conclusions

This study considers some model-based approaches that also contain NHIS design structures to complement some of the design-based approaches for analyzing NHIS-EPA linked data. As the nature of the linkage creates a possibility that the inferential population is not "representative" of the U.S., some adaptation of standard NHIS design-based methods may be required. First, by a model-based analysis we showed that the populations inferred by the linked and unlinked NHIS may have different characteristics, suggesting that care must be taken when making a population inference. Second, we considered some model-based approaches to association analyses between NHIS health variables and EPA pollutant variables. The simple models suggested making use of the survey design information, and subject to choosing reasonable conceptual models, allow the analyst to study associations while capturing much of the correct stochastic structure of the data. Caution is always advised as to the level of population inference.

References

Korn, E, Graubard B (1999). *Analysis of Health Surveys*: Wiley.

Judson DH, Parker JD, Larsen MD. (2013). Adjusting sample weights for linkage-eligibility using SUDAAN. National Center for Health Statistics, Hyattsville Maryland. Available at the following address: http://www.cdc.gov/nchs/data/datalinkage/adjusting_sample_weights_for_linkage_eligibility_using_sudaan.pdf

Parker JD, Kravets N, Woodruff TJ. (2008a). Linkage of the National Health Interview Survey to air quality data. National Center for Health Statistics. Vital Health Stat(145):1–24.

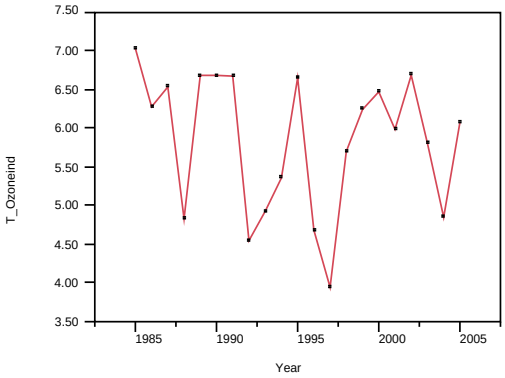
Parker DJ, Woodruff TJ, Akinbami LJ, Kravets N (2008). Linkage of the US National Health Interview Survey to air monitoring. Environmental Research 106 384–392.

Parker DJ, Woodruff TJ, Akinbami LJ (2009). Air Pollution and Childhood Respiratory Allergies in the United States. Environmental Health Perspectives volume 117 number 1 page 139-147.

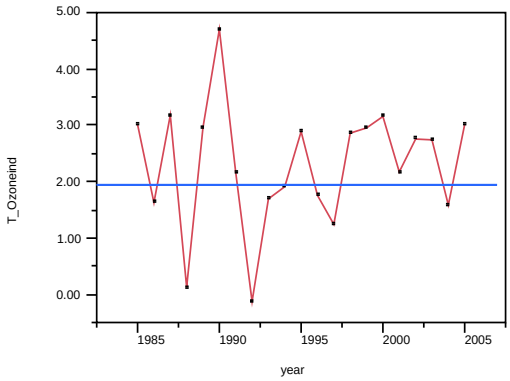
Wei R, and Parsons V, (2009). Model-based Methods in Analyzing Complex Survey Data: A Case Study with National Health Interview Survey data. ASA Proceedings of the Joint Statistical Meetings, pp 2558-2567.

Figure 1. T-statistic values for testing NHIS-EPA pollutant linkage domain effect for "unhealthy" status

1a. Single ozone indicator

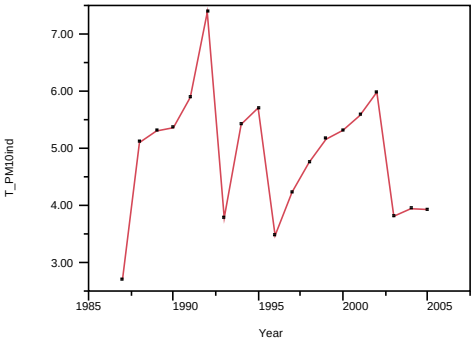


1b. Ozone indicator simultaneous with other indicators

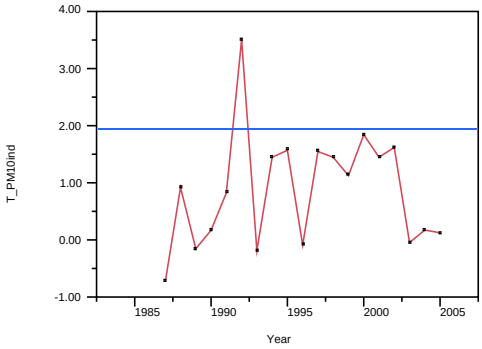


Horizontal line at 2.00 clarifies t-test significance region

1c. Single PM10 indicator



1d. PM10 indicator simultaneous with other indicators



Horizontal line at 2.00 clarifies t-test significance region

Figure 2. Gender effect estimates for "unhealthy" status for complete NHIS data (red) and partial NHIS-EPA linked data (blue)

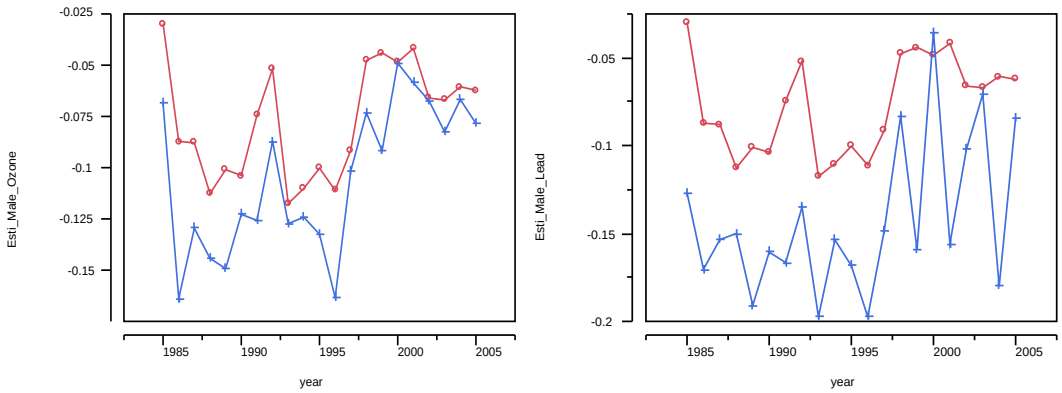
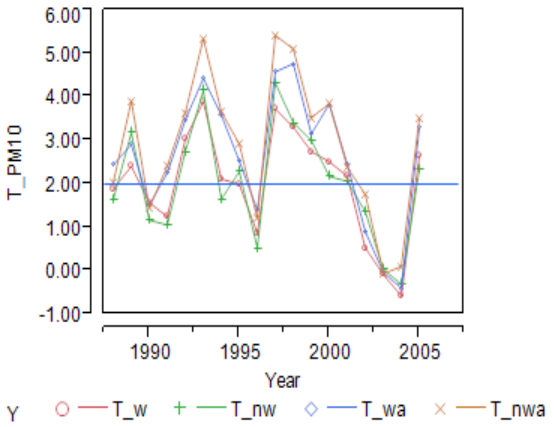
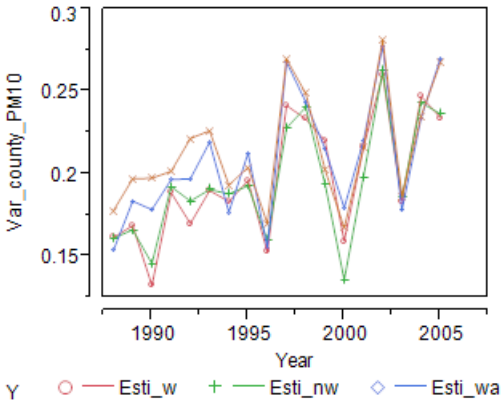


Figure 3. Comparisons of four models that express associations between "unhealthy" status and level of pollutant:
T-statistics for testing significance of the fixed pollutant effect α,
Random effect variance estimates, σ^2_{cnty},

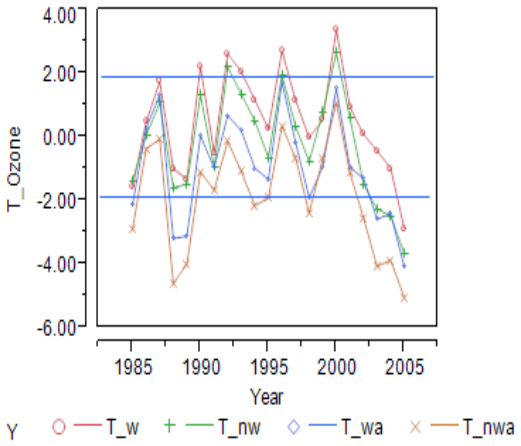
3a T-statistic PM10



3b Random effect variance PM10



3c T-statistic Ozone



3d Random effect variance Ozone

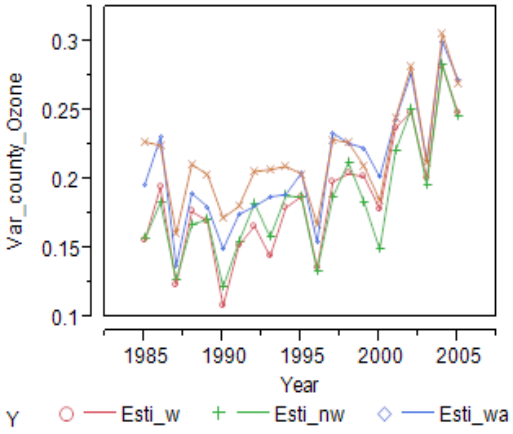
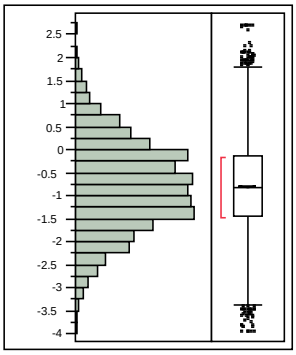


Figure 4. Bayesian Posterior Distributions

a. Ozone



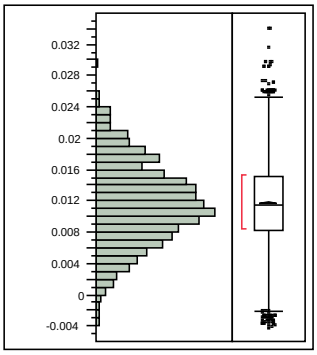
Quantiles

100.0%	maximum	2.68788
99.5%		1.76693
97.5%		1.17166
90.0%		0.45329
75.0%	quartile	-0.1405
50.0%	median	-0.8122
25.0%	quartile	-1.44
10.0%		-2.0766
2.5%		-2.7039
0.5%		-3.2856
0.0%	minimum	-3.9549

Summary Statistics

Mean	-0.802537
Std Dev	0.9768013
Std Err Mean	0.009768
Upper 95% Mean	-0.78339
Lower 95% Mean	-0.821684
N	10000

b. PM10



Quantiles

100.0%	maximum	0.03396
99.5%		0.02495
7.5%		0.02236
90.0%		0.01853
75.0%	quartile	0.01501
50.0%	median	0.01147
25.0%	quartile	0.00816
10.0%		0.00502
2.5%		0.00169
0.5%		-0.0019
0.0%	minimum	-0.0043

Summary Statistics

Mean	0.0116703
Std Dev	0.0052191
Std Err Mean	0.0000522
Upper 95% Mean	0.0117727
Lower 95% Mean	0.011568
N	10000

Table 2. Comparisons of SAS GLIMMIX and MCMC procedures for modeling associations between "unhealthy" status and pollutant level

	NHIS[1] full	Ozone		PM2.5		PM10		CO1		NO2		Lead	
		GLIMMIX	MCMC	GLIMMIX	MCMC	GLIMMIX	MCMC	GLIMMIX	MCMC	GLIMMIX	MCMC	GLIMMIX	MCMC
race(hisp)[2]	0.681	0.759	0.779	0.757	0.752	0.813	0.803	0.821	0.804	0.811	0.801	0.890	0.832
race(other)	0.078	0.046	0.066	0.074	0.074	0.129	0.108	0.117	0.065	0.107	0.102	0.256	0.087
race(black)	0.735	0.753	0.755	0.714	0.705	0.754	0.745	0.758	0.744	0.777	0.762	0.789	0.748
sex(male)[3]	-0.062	-0.096	-0.098	-0.103	-0.105	-0.126	-0.128	-0.104	-0.105	-0.109	-0.109	-0.108	-0.108
age45	0.047	0.048	0.049	0.048	0.049	0.048	0.049	0.049	0.049	0.049	0.049	0.048	0.045
air pollutant		-14.141	-0.803	0.038	0.060	0.010	0.012	0.033	0.218	3.553	0.502	-0.674	-0.669

1/ for complete data
2/ white is the reference group
3/ female is the reference group
Green numbers are not significant at the 0.05 level.