2008

Linkage of the US National Health Interview Survey to air monitoring data: An evaluation of different strategies

Jennifer D. Parker
National Center for Health Statistics, CDC, 3311 Toledo Road, Room 6107, Hyattsville, MD 20782, USA

Tracey Woodruff
US Environmental Protection Agency, University of California, San Francisco, USA

Lara J. Akinbami
National Center for Health Statistics, CDC, 3311 Toledo Road, Room 6107, Hyattsville, MD 20782, USA

Natalyia Kravets
NOVA Research Company, Northrop Grumman CITS II Contract, USA

Available online at www.sciencedirect.com
ScienceDirect
Environmental Research 106 (2008) 384–392
Environmental Research
www.elsevier.com/locate/envres

Linkage of the US National Health Interview Survey to air monitoring data: An evaluation of different strategies ☆

Jennifer D. Parker[a,*], Tracey J. Woodruff[b], Lara J. Akinbami[a], Nataliya Kravets[c]

[a]National Center for Health Statistics, CDC, 3311 Toledo Road, Room 6107, Hyattsville, MD 20782, USA
[b]US Environmental Protection Agency, University of California, San Francisco, USA
[c]NOVA Research Company, Northrop Grumman CITS II Contract, USA

Received 1 June 2007; received in revised form 14 September 2007; accepted 1 November 2007
Available online 20 February 2008

Abstract

The goal of this study is to describe linkages between the National Health Interview Survey (NHIS) and Environmental Protection Agency (EPA) air monitoring data, specifically how the linkage method affects characteristics and exposure estimates of study samples and estimated associations between exposure and health. In the USA, nationally representative health data are collected in the NHIS and annual air quality data are collected by the EPA. The linkage of these data for research is not straightforward and the choices made may introduce bias into results. The 2000–2003 NHIS and air quality data for six air pollutants were linked by residential block group and monitor location, which differ by pollutants. For each pollutant, three annual exposure variables were assigned to respondents: (1) average of all monitors in the county, (2) of monitors within a 5-mile radius of the distance between block group and monitor, and (3) within a 20-mile radius. Exposure estimates, study sample characteristics, and association between fine particle exposure and respondent-reported health status were compared for different geographic linkage methods. The results showed that study sample characteristics varied by geographic linkage method and pollutant linked. Generally, the fewer the NHIS respondents linked, the higher is the pollution exposure and lower is the percentage of non-Hispanic whites. After adjustment for sociodemographic and geographic factors, associations between fine particles and health status were generally comparable across study samples. Because exposure information is not available for all potential participants in an epidemiological study, selection effects should be considered when drawing inferences about air quality–health associations. With the current monitoring data system, the study sample is substantially reduced when linkage to multiple pollutants is performed.

Keywords: Air pollution; Health status; Selection bias; National Health Interview Survey; Fine particles

1. Introduction

Adverse health effects of ambient exposure to environmental contaminants have been convincingly demonstrated in epidemiological research. However, studies in the United States examining pollution and health outcomes have been conducted using select populations in limited geographic areas. The degree to which the results may reflect the experience of more general populations is unknown. In the USA, data on environmental contaminants are collected by the US Environmental Protection Agency (US EPA) and nationally representative health data are collected by the National Center for Health Statistics/Centers for Disease Control and Prevention (NCHS/CDC). Opportunities exist to link these data to obtain a more nationally representative study sample than possible with geographically limited data.

Only a few previous studies have combined NCHS/CDC survey data with US EPA exposure data, likely due to confidentiality restrictions and the resulting unavailability of geographic details on public-use files. Ostro combined data from the National Health Interview Survey (NHIS)

☆The National Health Interview Survey is an ongoing national survey in the United States. No additional human subjects approval was required for this linkage study.

*Corresponding author. Fax: +1 301 458 3078.
E-mail address: jdparker@cdc.gov (J.D. Parker).

0013-9351/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.envres.2007.11.001

from the late 1970s with corresponding estimates of air pollution to examine respiratory morbidity and work-loss days (Ostro, 1983, 1987, 1989, 1990; Ostro and Rothschild, 1989). Later, data from the National Health and Nutrition Examination Surveys (NHANES) were combined with exposure estimates from the US EPA to examine the effects of air pollution on lung function and its correlation with blood markers (Chestnut et al., 1991; Schwartz, 1989, 2001). In a methodological study, Wong et al. (2004) compared pollution exposure estimates for children in NHANES III using four different assignment methods. In addition to survey data, vital statistics, compiled at NCHS/CDC, have been linked to air pollution data in several studies (Bell et al., 2004; Darrow et al., 2006; Dominici et al., 2002; Samet et al., 2000; Woodruff et al., 1997).

In the examples listed above, individual-level outcomes and characteristics are available while the air pollution exposure is aggregated (Kunzli and Tager, 1997). In environmental studies, the aggregated exposure is often assumed to be an approximate exposure assignment for an individual. From this perspective, the exposure variable is considered subject to measurement error, and resulting associations are often assumed to be attenuated, although this assumption may not hold (Budtz-Jorgensen et al., 2003; Greenland and Gustafson, 2006; Zeger et al., 2000). A related issue when creating area-level averages is the Modifiable Unit of Analysis Problem, where associations differ by the level of aggregation. Differences are due to both mathematical properties of aggregation and specification bias, where characteristics of groups differ by the aggregation (Waller and Gotway, 2004).

In studies based on linked NCHS survey data and US EPA data cited above, exposure estimates assigned to survey respondents were derived in a variety of ways, including averages over metropolitan areas (Ostro, 1983, 1987, 1989, 1990; Ostro and Rothschild, 1989; Woodruff et al., 1997), or counties (Darrow et al., 2006; Schwartz, 2001) and using monitors within specified distances from the respondent's residence (Schwartz, 1989). Choice of geographic scale for assigning environmental exposures has been compared in only a few studies. Willis et al. (2003) conducted a re-analysis of the American Cancer Society Study (ACS) and found stronger associations using exposures calculated at the county level than at the original metropolitan area level. In contrast, Basu et al. (2004) compared county-level pollution exposure to exposure based on averaging pollution measurements within 5 miles of a mother's residence to assess $PM_{2.5}$ exposure and birth weight; the results of this California study showed a stronger effect for the county level compared to the 5-mile exposure measure. The consequences of using different geographic units of analysis have also been compared in studies of neighborhood characteristics on health (Krieger et al., 2002).

An issue that has not been fully examined in environmental epidemiology is selection bias. Selection bias occurs when the relationship between the outcome and exposure for subjects included in the analysis is different from the relationship for those not included (Ellenberg, 1994). In the linked studies using NCHS datasets described above, the analytic samples excluded varying percentages of the surveyed responders due to insufficient exposure information. In the studies by Ostro, for example, the findings are based on approximately 7000–8000 NHIS working adults with exposure information per survey year; using the 1979 NHIS, we calculated that nearly 45,000 of the respondents were working adults, indicating that exposure information was not available for most of the eligible survey respondents (data available at http://www.cdc.gov/nchs/about/major/nhis/quest_data_related_1969_96.htm). In the re-analysis of the ACS described above, fewer than half of the study subjects in the original metropolitan-area-level analysis were available for the county-level analysis (Willis et al., 2003). The study by Basu et al. (2004), on the other hand, used the same study cohort for both the county-level and 5-mile analyses. Whether the conclusions of Willis et al. would have differed had all respondents been included in the county-level analysis is unknown.

The objective of this paper is to compare the study samples that result from using different linkage approaches that vary by geographic scale and number of air pollutants when combining the NHIS with air pollution data from the US EPA. Different geographic linkage decisions lead both to different study samples and to different exposure assignments, either of which can lead to varying results. An understanding of the effects of geographic linkage decisions on the characteristics of the study sample and exposure assignment is needed to further understand whether and how pollution and health relate to each other. This evaluation has implications for studies of chronic or long-term exposure to air pollution; time-series studies of daily events may be less affected by linkage issues.

Toward this objective, we linked respondents in the 2000–2003 NHIS to annual monitoring averages for six criteria pollutants: particulates, fine ($PM_{2.5}$) and large (PM_{10}); carbon monoxide (CO); sulfur dioxide (SO_2); ozone (O_3); and nitrogen dioxide (NO_2). Respondents were linked, when possible, to monitor data in their county of residence, to data from a monitor within 5 miles of their block group, and to data from a monitor within 20 miles of their block group separately for each pollutant. To examine the effect on subsequent inference of requiring linkage to multiple pollutants, additional comparisons were made for study samples defined by residential linkage to all six pollutants. Because of recent interest in the health effects of fine particles (Pope and Dockery, 2006), demographic and health characteristics were compared for study samples linked to $PM_{2.5}$ exposures as an example.

Using a general health indicator (fair/poor versus good/very good/excellent respondent-reported health status), associations between exposure and health status were evaluated. Because a thorough examination of the effects of air pollution on a particular health outcome was not intended, we used a general measure of health as an

indicator of the underlying health of the population, rather than a health outcome specific to air pollution, to examine the effects of geographic linkage decisions on selection bias and exposure assignment in national health data.

2. Materials and methods

The NHIS is a survey of a nationally representative sample of the civilian, non-institutionalized population conducted continuously by NCHS/CDC. Sociodemographic information and answers to a variety of health-related questions are obtained for each household member and included in the NHIS Person file (Schille et al., 2005). In addition, more detailed health-related questions are asked to a randomly selected sample adult from each family and a randomly selected sample child from each family with children and are contained in the NHIS Sample Adult and NHIS Sample Child files (Dey and Bloom, 2005; Lethbridge-Cejku and Vickerie, 2005).

For this analysis, geocoded NHIS data from the 2000–2003 survey were used. Files containing geographic detail are not available publicly but can be used for research purposes through the NCHS Research Data Center (RDC). Although more recent NHIS data years are available, these files were not geocoded at the time of this study. Geographic variables used to assign pollution exposure for each respondent were county and latitude and longitude of the population center of the census block group of residence. Population centers are location indicators weighted by population size. A recent paper by Kravets and Hadden (2007) provides more detail on geocoding the NHIS. Although the geocoding varies by year, for the 2000–2003 survey, nearly all (99.9%) respondents could be geocoded to a Census 2000 block group. The exact residential locations are not retained in the final analytic or in-house files.

Race and Hispanic origin were combined into seven distinct categories: white, black, Asian, Hispanic, American Indian and Alaska Native (AIAN), multiple race, and other race. Race and Hispanic origin are collected separately in the NHIS; for this analysis, any respondent reporting Hispanic origin was assigned the Hispanic category, regardless of race. For the Hispanic respondents, those reporting Mexican or Mexican American origin were categorized as a subgroup. Family income refers to the total family income received in the previous calendar year by all family members. Income is converted into a percent of the official poverty threshold by taking into account both the total family income and family size. Responses were grouped into four levels of income as a percent of poverty: less than 100%, 100–199%, 200–399%, and 400% or more. Because the number of respondents missing family income is relatively large, we used the data files produced by NCHS with missing values of family income imputed using multiple imputation methods (Schenker et al., 2006). Race/ethnicity and family income are both strongly associated with health outcomes (NCHS, 2007). Furthermore, because disadvantaged groups are more likely to live in areas with higher levels of pollution (Lee, 2002), these sociodemographic factors may affect exposure–health associations. Geographic descriptors were census region and the 2006 Urban–Rural Continuum (Ingram and Franco, 2006); pollution exposure, sociodemographic characteristics, and health outcomes differ by urban/rural status (Eberhardt and Pamuk, 2004; US EPA, 2003).

As potential outcome and confounding variables in environmental health studies, a handful of health variables were included in the tabulations. The percentage reporting fair/poor health status was calculated for all respondents. For the sample adults, we tabulated cigarette smoking (current, former, or never). For the sample adults and sample children, the NHIS collects information about a number of conditions. We tabulated the percentages that had been told by a doctor or health professional in the previous 12 months that they had chronic bronchitis (adults), sinusitis (adults), respiratory infection (children), ear infection (children), and hay fever (adults and children). These health indicators were not intended to be an exhaustive set of potential pollution-associated measures in the NHIS.

Annual averages of pollution exposure by pollution monitor for six air pollutants for 2000–2003 were obtained from the US EPA's Air Quality System (AQS) database (US EPA, 2006). The AQS provides air monitoring data-ambient concentrations of criteria and hazardous air pollutants at monitoring sites throughout the USA, primarily in cities and towns; these data are collected for regulatory purposes. The pollutants $PM_{2.5}$, PM_{10}, CO, SO_2, O_3, and NO_2 are referred to as criteria pollutants and are routinely monitored and regulated by the US EPA. Studies have demonstrated that higher ambient levels of these pollutants are correlated with poorer health outcomes (US EPA, 2003). Air quality measurements are obtained at different time intervals. Some monitors collect information daily, for example, while others collect information every third or sixth day. Monitors were included in this study if data were available for at least 75% of the scheduled monitoring times. A small number of monitors had county location but not latitude and longitude and were linked by county but not by distance. Additional details on the selection of monitors for this analysis are available from the authors.

An average annual exposure of each pollutant, if available, was assigned to each NHIS respondent by averaging the annual exposures from all monitors within the county and all monitors within 5 and 20 miles of the block group. For the 5-mile and 20-mile exposures, weighted averages were calculated using the inverse of the squared distance between the block group center and the monitor as the weight.

Three types of linkages for each pollutant were defined based on the following geographic criteria and the availability of data: (1) respondents linked to monitoring data by county, (2) respondents linked to monitors within 20 miles, and (3) respondents linked to monitors within 5 miles. Using these three geographic criteria, 18 separate geographic study samples were formed by linking to each of the six pollutants and an additional three were formed by simultaneously linking to all six pollutants. For each of 21 geographic study samples, the median and interquartile ranges (25th and 75th percentiles) of the annual average pollution values were calculated.

Focusing on the three linkages to $PM_{2.5}$ exposure (county-level, the 20-mile, and the 5-mile) and the three linkages to all six measured pollutants, we characterized the resulting geographic study samples by the socio-demographic, geographic, and health indicators described above.

The association between respondent-reported health status, defined as a dichotomous variable "excellent/very good/good" versus "fair/poor," and $PM_{2.5}$ exposure was estimated for each of the subgroups using standard logistic regression analysis for surveys with SUDAAN (RTI International, 2006), including appropriate methods for the multiple imputations of poverty status (Schenker et al., 2006). The associations for a $10\,\mu g/m^3$ change in $PM_{2.5}$ are presented unadjusted as well as adjusted for socioeconomic and geographic variables. Geographic variation in health and pollution levels exists but its role in these models is uncertain. If, for example, geographic variation in health can be attributed to air pollution, then controlling for region of the country may mask the effect. Alternatively, exposure measures may be more representative for urban compared to suburban areas. Although a thorough understanding of potential regional variation in the effects of air pollution on health is beyond the scope of this descriptive analysis, models were fit both with and without the geographic variables but exploration of potential geographic effect modification was not done. Similarly, while socio-economic variables are established confounders, their role as effect modifiers is uncertain. On one hand, disadvantaged groups may be more susceptible to the effects of air pollutants; on the other hand, other factors that contribute to poor health outcomes may overwhelm additional effects of pollution for these groups. A full exploration of effect modification by socioeconomic status was beyond the scope of this analysis.

Because smoking is a strong determinant of health outcomes and smoking status varies across the USA, smoking could be considered a strong confounder of pollution–health associations. To assess the impact of smoking status as a confounder on these associations, using a sub-sample of data from the Sample Adult file (45% of adults), models were fit with and without the indicator of smoking and for the subset of never smokers.

To separate the effect of the study sample from the effect of the exposure assignment, models were fit to estimate associations between health status and $PM_{2.5}$ using a study sample with values for $PM_{2.5}$ defined

Table 1
Number and weighted percentage of NHIS respondents linked to annual average exposure data for six criteria pollutants by geographic linkage method, 2000–2003

Pollutant	Geographic linkage method					
	Monitors in county of residence		Monitors within 5 miles of residence		Monitors within 20 miles of residence	
	N	%	*N*	%	*N*	%
$PM_{2.5}$	282,935	71	155,045	37	313,047	80
PM_{10}	221,160	53	118,294	27	274,126	68
CO	208,359	49	109,117	24	246,755	61
O_3	273,506	69	135,921	31	308,960	79
NO_2	187,769	44	89,979	19	226,335	55
SO_2	186,103	45	76,461	18	230,294	58
All six	125,284	28	30,461	6	176,546	42

Table 2
Median and the interquartile range (25th and 75th percentiles) of the annual average pollution exposure for NHIS survey respondents linked to exposure data for single pollutants by geographic linkage method, 2000–2003

Pollutant	Geographic linkage method[a]					
	Monitors in county of residence		Monitors within 5 miles of residence		Monitors within 20 miles of residence	
	Median	25th–75th percentiles	Median	25th–75th percentiles	Median	25th–75th percentiles
$PM_{2.5}$ ($\mu g/m^3$)	13	11.0–15.3	13.5	11.3–15.5	13	11–15.1
PM_{10} ($\mu g/m^3$)	24.0	20.5–29.6	24.4	20.8–29.6	24.1	21.1–28.1
CO (ppm × 100)	60	47–82	64	49–87	61	48–81
O_3 (ppm × 1000)	52	47–57	51	46–56	52	48–57
NO_2 (ppm × 1000)	18	13–24	20	16–24	18	14–27
SO_2 (ppm × 10,000)	37	21–52	41	23–61	45	27–57

[a]See Table 1 for underlying sample sizes.

at the county level and for both 5 and 20 miles and a study sample with $PM_{2.5}$ and the other pollutants defined at all levels. The study sample linked to all three geographic scenarios for $PM_{2.5}$ is slightly smaller than the 5-mile study sample because some respondents in the 5-mile study sample do not have a county-level measure, that is, the monitor (or monitors) within 5 miles is in a different county.

For comparisons across study samples, we inspected the estimates but statistical tests were not done. The study samples overlap and are, thus, not independent. Because of the complex survey design of the NHIS, standard errors for all estimates were calculated using standard survey methods with the software SUDAAN (RTI International, 2006) and all percentages and odds ratios (ORs) are weighted to represent the civilian non-institutionalized population; sample sizes are unweighted.

3. Results

There were over 380,000 respondents in the 2000–2003 NHIS. The percentage of survey respondents linked to pollution data varied by pollutant and geographic linkage (Table 1). For all pollutants, the percentage of respondents linked is slightly higher using the 20-mile radius than using a county linkage, due, in part, to some county boundaries being less than 20 miles from a respondent's location. Linkage to all six pollutants decreases the percentage of respondents available for analysis markedly, particularly using 5-mile radii.

Average annual exposure measures are relatively similar across linkage methods (Table 2). Exposure estimates tend to be slightly higher for the 5-mile linkage. Among the respondents linked to six pollutants, exposure estimates for particulate matter were higher than those linked to just one pollutant (Table 3); this effect was less pronounced for gaseous pollutants.

Using a more restricted sample of approximately 25,000 respondents with exposure estimates for all six pollutants and all three geographic linkage methods, pollution-specific correlations for the different pollution-specific exposure measures at different geographic scales were all greater than 0.85 (not shown); using $PM_{2.5}$, as an example, the correlations between the county-level estimate of $PM_{2.5}$ and those calculated for 5 and 20 miles were 0.91 and 0.93, respectively. Given that many of the same monitors are used in the calculation of these exposures for a respondent, the high correlations are not surprising. Nevertheless, the high correlations at the individual level suggest that, on average, the exposure estimates assigned using the 5-mile criterion are very similar to those assigned at the county level.

Demographic distributions differed among the study samples linked to $PM_{2.5}$ data (Table 4). The percentage of

Table 3
Median and the interquartile range (25th and 75th percentiles) for annual average pollution exposure for NHIS survey respondents linked to exposure data for all six pollutants by geographic linkage method, 2000–2003

Pollutant	Geographic linkage method					
	Monitors in county of residence ($N = 125{,}284$)		Monitors within 5 miles of residence ($N = 30{,}461$)		Monitors within 20 miles of residence ($N = 176{,}546$)	
	Median	25th–75th percentiles	Median	25th–75th percentiles	Median	25th–75th percentiles
$PM_{2.5}$ ($\mu g/m^3$)	14.3	11.4–16.6	15.2	13.3–16.2	14.3	11.9–15.7
PM_{10} ($\mu g/m^3$)	26.4	22.3–32.6	26.0	22.2–34.2	25.1	22.2–29.9
CO (ppm × 100)	65	52–84	70	54–93	65	53–84
O_3 (ppm × 1000)	51	47–55	49	43–53	52	47–56
NO_2 (ppm × 1000)	19	14–25	23	19–30	19	15–26
SO_2 (ppm × 10,000)	30	21–46	46	25–64	40	22–57

white respondents, for example, was highest in the study sample linked to $PM_{2.5}$ monitors within 20 miles and lowest in the study sample linked to $PM_{2.5}$ monitors within 5 miles. The percentage of respondents in poverty was highest in the study sample linked to $PM_{2.5}$ monitors within 5 miles. The distributions of region and level of urbanicity differed among the study samples. The prevalence of the health conditions reported for the sample adult or child from the previous 12 months were, with one exception, higher for the overall NHIS than for any of the linkage study samples defined by geographic linkage to $PM_{2.5}$. The age distribution appears similar across study samples.

The results for the study samples linked to all six pollutants were similar to those for the study samples defined by linkage to $PM_{2.5}$. For example, the study samples linked to monitors for all six pollutants within 5 miles had fewer white respondents, more Hispanic respondents, and respondents from families with lower incomes than either study samples defined by county linkages or linked to multiple pollutants within 20 miles (Table 4). As above, distributions of age and health outcomes were similar, though the 5-mile sample included a slightly higher percentage of smokers and a lower percentage of young children with ear infections; the prevalence of the health conditions reported for the sample adult or child from the previous 12 months were generally higher for the overall NHIS than for any of the linkage study samples. Again, there were large differences by region and urbanicity. Compared to the overall NHIS, for example, respondents from the West are over represented after linking to multiple pollutants by county.

The odds of reporting fair or poor health status increased with an increase of $PM_{2.5}$ of $10\,\mu g/m^3$; this increase was approximately 10–20% but varied depending on the study sample and adjustment for covariates (Table 5). Associations were attenuated with adjustment for sociodemographic and geographic variables. The effect of the adjustment was somewhat more pronounced for the 5-mile linkage subgroups than for the other groups.

Among the subset of adults with smoking information, ORs were similar to those reported in Table 5 (not shown). While smoking status consistently had an independent effect on health status, its inclusion in the regression models did not modify any association by more than 1% (not shown). Among the subset of never smokers, all associations between health status and air pollution were stronger and the variation in effect estimates among linkage study samples was similar to the overall results (not shown).

To assess whether variability in findings from different geographic linkage methods can be attributed to differences in exposure assignments or differences in study samples, we compared associations between health status and $PM_{2.5}$ using the same respondents. For this analysis, we used two study samples of respondents, each with county-level, 20-mile, and 5-mile $PM_{2.5}$ exposure measures. The first subset required only $PM_{2.5}$ exposures for each geographic level and the second subset comprised the approximately 25,000 respondents with linkage to all pollutants at each geographic level; the sample size differs from that for 5 miles because some respondents with 5-mile exposures did not have county-level exposure estimates. Adjusted ORs using the county-level and 20-mile exposure variables were the same as those calculated using the 5-mile exposure variable despite the wider geographic areas used in their definition (Table 6). This finding is consistent with the high correlations reported above. This similarity suggests that some of the differences in ORs observed in Table 5 may be due to differences in study sample rather than the exposure variable definition.

4. Discussion

These findings show variation in analytic study samples derived from a single nationally representative database (the NHIS) when different approaches are used to link the NHIS to EPA air pollution data. The variation among study samples is most pronounced for demographic and geographic (region and urbanicity) variables. In general,

Table 4
Characteristics of NHIS respondents linked to $PM_{2.5}$ monitors by geographic linkage method (percentage), 2000–2003

	All	Linked to $PM_{2.5}$			Linked to all six pollutants		
		County	5 miles	20 miles	County	5 miles	20 miles
Sample size	383,995	282,935	155,045	313,047	125,284	30,461	176,546
Percentage							
Race/ethnicity							
Asian	3.6	4.6	5.2	4.3	5.8	7.2	5.6
American Indian/Alaska Native	0.6	0.5	0.4	0.4	0.3	0.3	0.3
Black	12.0	14.0	17.4	13.3	15.9	22.0	16.9
Hispanic	12.8	15.6	17.4	14.0	24.2	29.9	18.7
Mexican	8.3	9.8	10.6	8.7	15.6	16.1	10.6
Other Hispanic	4.5	5.8	6.8	5.3	8.7	13.8	8.1
White	69.6	63.9	58.2	66.6	52.1	39.3	57.2
Two or more races	0.9	0.9	0.9	0.9	1.0	0.8	0.8
Other	0.5	0.5	0.6	0.5	0.6	0.6	0.5
Poverty level							
<100%	13.2	13.1	15.9	12.7	14.5	22.3	13.2
100–200%	18.7	18.1	19.7	17.9	19.4	22.6	17.8
200–400%	31.4	30.6	30.9	30.7	29.6	27.7	29.5
>400%	36.7	38.2	33.6	38.8	36.5	27.4	39.4
Age (years)							
<17	24.7	25.0	24.6	24.9	25.5	25.9	25.0
17–44	40.5	41.2	42.2	41.2	42.0	43.6	42.0
45–64	22.9	22.4	21.5	22.6	21.5	19.7	21.9
>64	11.9	11.5	11.7	11.4	11.0	10.8	11.4
Region							
Northwest	19.0	19.7	22.7	20.4	14.2	36.3	26.1
Midwest	24.0	21.6	23.9	21.7	17.1	14.8	18.2
South	36.2	32.5	28.3	34.7	29.9	16.5	29.9
West	20.8	26.3	25.2	23.2	38.8	32.5	25.8
Urban/rural county							
Large central metropolitan	28.5	40.0	43.8	35.1	73.0	73.1	54.8
Large fringe metropolitan	24.7	24.4	21.2	27.7	14.0	14.0	32.8
Medium metro	20.3	22.9	22.3	22.4	12.4	11.5	11.1
Small metro	9.8	6.9	8.4	7.6	0.6	0.4	0.6
Micro-politan	10.4	5.0	3.8	5.8	0	0	0.7
Non-core	6.2	0.8	0.5	1.5	0	0	0
Health characteristics							
Fair/poor health status	9.1	8.6	9.4	8.6	8.6	10.3	8.3
Smoking[a], adults							
Current	22.5	21.6	22.3	21.8	20.4	23.0	20.5
Former	22.2	21.8	20.6	22.0	20.0	17.4	20.8
Never	55.3	56.6	56.8	56.2	59.4	59.5	58.8
Health conditions diagnosed in past 12 months[a]							
Adults[a]							
Chronic bronchitis	4.6	4.3	4.4	4.5	3.9	3.9	3.9
Sinusitis	15.4	14.8	14.4	15.2	13.2	12.7	12.9
Hay fever	9.2	9.4	9.4	9.4	8.9	9.5	9.2
Children[a]							
Respiratory infection	13.0	12.4	11.7	12.8	10.5	9.3	11.1
Hay fever	11.3	11.3	10.8	11.4	9.7	10.1	10.6
Ear infection, 0–2 years	13.3	12.6	11.6	12.9	10.9	7.2	11.3
Ear infection, 3–17 years	5.2	5.0	4.8	5.1	4.5	4.3	4.7

[a]Includes only the sample adults or sample children asked detailed health questions; the few respondents with missing data for a particular health question were excluded.

Table 5
Unadjusted and adjusted odds ratios (OR) with 95% confidence intervals (CI) describing association between respondent-reported fair/poor health status and $PM_{2.5}$ exposure (per 10 μg/m^3) by geographic linkage method

	Linked to $PM_{2.5}$			Linked to all six pollutants[a]		
	County	5 miles	20 miles	County	5 miles	20 miles
Sample size	282,935	155,045	313,047	125,284	30,461	176,546
Unadjusted						
OR	1.22	1.25	1.15	1.27	1.18	1.22
95% CI	1.14–1.31	1.17–1.34	1.09–1.22	1.18–1.37	1.03–1.35	1.14–1.30
Adjusted						
OR[b]	1.15	1.10	1.11	1.26	1.09	1.16
95% CI	1.08–1.22	1.02–1.17	1.06–1.17	1.16–1.36	0.95–1.26	1.09–1.24
OR[c]	1.22	1.14	1.18	1.24	1.10	1.16
95% CI	1.15–1.30	1.07–1.22	1.12–1.25	1.12–1.37	0.95–1.28	1.07–1.25

[a]Linked to $PM_{2.5}$, PM_{10}, carbon monoxide, sulfur dioxide, ozone, and nitrogen dioxide exposure.
[b]Adjusted for race and ethnicity, poverty status, and age.
[c]Adjusted for race and ethnicity, poverty status, age, region, and urbanicity.

Table 6
Unadjusted and adjusted odds ratios (OR) with 95% confidence intervals (CI) describing association between respondent-reported fair/poor health status and $PM_{2.5}$ exposure (per 10 μg/m^3) among respondents with $PM_{2.5}$ linkage for all geographic linkage methods (5 miles, 20 miles, and county level)

	Linked to $PM_{2.5}$			Linked to multiple pollutants[a]		
	County	5 miles	20 miles	County	5 miles	20 miles
Sample size	151,870	151,870	151,870	25,687	25,687	25,687
Unadjusted						
OR	1.27	1.20	1.25	1.17	1.14	1.11
95% CI	1.17–1.38	1.12–1.28	1.17–1.34	0.99–1.40	1.00–1.31	0.97–1.27
Adjusted						
OR[b]	1.10	1.08	1.10	1.08	1.09	1.08
95% CI	1.02–1.19	1.01–1.58	1.02–1.17	0.91–1.30	0.95–1.25	0.93–1.24
OR[c]	1.15	1.13	1.14	1.09	1.10	1.09
95% CI	1.06–1.24	1.06–1.21	1.07–1.22	0.89–1.32	0.94–1.28	0.93–1.28

[a]Linked to $PM_{2.5}$, PM_{10}, carbon monoxide, sulfur dioxide, ozone, and nitrogen dioxide exposure. The sample sizes are smaller than the previous 5-mile sample sizes (Tables 1–5) because some respondents with 5-mile exposure estimates do not have county-level exposure estimates.
[b]Adjusted for race and ethnicity, poverty status, and age.
[c]Adjusted for race and ethnicity, poverty status, age, region, and urbanicity.

the more restrictive the linkage criteria, the more urban the resulting study sample becomes. Correspondingly, the underlying study samples differed in important ways in a variety of demographic factors. In an analysis using data linked by county to multiple pollutants, the resulting study population would be considerably smaller than the original NHIS and have nearly twice the percentage of respondents from the West. There may be trade-offs between a seemingly more precise geographic area using a smaller radius and the loss of statistical precision with fewer survey respondents. Although the NHIS is a large survey, the number of respondents with uncommon health outcomes or in subpopulations can be small. Measurement error and other issues that plague single pollution studies are already exacerbated in analyses of multiple pollutants (Zeka and Schwartz, 2004); the effects of selection bias and reduced sample size with linked multiple pollutant data add to the challenges of analysis and interpretation.

Associations between health outcomes and air pollution can depend both on the underlying study sample and the geographic exposure assignment. We found that differences in the geographic linkage method did not lead to large differences in the association between health status and $PM_{2.5}$ pollution. That the adjusted ORs calculated for the 5-mile study samples were somewhat smaller than those calculated using the broader exposure areas (Table 5) is consistent with the findings of Basu et al. (2004) and does not support the hypothesis that more precise measurements lead to stronger associations. However, it is unclear whether the results in Table 5 are due to differences in study sample or differences in exposure assignment. It is possible that the effect of sample selection (or specification

bias) is greater than the effect of exposure assignment (or measurement error), but numerous other factors could also be contributing to these differences. For example, in addition to population differences, measurement error likely differs among areas defined by urban–rural status; exposure for respondents in more suburban areas may be less well characterized than those in more urban areas. The composition of the fine particles near to monitors may differ from that farther away which, in turn, could lead to different observed health effects; additionally, some monitors may be cited near areas at risk of non-compliance potentially leading to population and exposure differences. The results in Table 6 address this question, in part, by using the same study samples with different exposure assignments. That the adjusted ORs for the county-level and 20-mile exposures are similar to the 5-mile exposure using the common study samples supports a greater impact of sample selection than of measurement error on the results (Table 6). This suggests that county-level measurements may provide reasonable estimates of fine particle exposure for some outcomes in air pollution studies, creating more opportunities for research.

The results from the models were not intended to provide conclusive evidence on the possible effects of fine particulate matter on general health status, rather they were intended to illustrate potential effects of sample selection and exposure assignment due to monitor locations on inferences. Yet, while the general association between $PM_{2.5}$ and health status varied somewhat by analytic approach, the results generally suggest a 10–20% increased odds of fair or poor health status with a $10\,\mu g/m^3$ change in $PM_{2.5}$ after controlling for potential confounders. Despite the imprecision of general health status in an interview survey, health status has been shown to be related to mortality even after controlling for known social, demographic, and medical factors; one recent study found particularly strong associations between health status and mortality due to respiratory and infectious causes and diabetes (Benjamins et al., 2004). Within our data, a full exploration of the relationship between health status and other health conditions, considering demographic and co-existing health measures, was not done. Nevertheless, our reported findings support further examination of the effects of air pollution on health using more specific health outcomes.

A limitation common to environmental health studies is the approximation of exposure based on residential location, or in this study, block group location, rather than personal exposures. Additionally, the use of annual averages rather than more targeted time intervals provides a relative ranking of air quality by location; but may not provide enough information for studying some health effects. The ability to assign exposure more precisely would affect associations reported for health status and $PM_{2.5}$, but would not alter the conclusion that findings using linked national data systems may be subject to selection effects.

With the increasing ability to assign exposure for study participants in areas without actual monitors using spatial models (Jerrett et al., 2005) or multiple imputation methods (Le et al., 2006; Sheppard et al., 1999), the necessity for study subjects to be situated near monitors may lessen. However, more work is needed to ascertain whether the modeled data generally performs similarly as measured data in epidemiological studies. Importantly, for national data systems, the accuracy of modeled data, as the accuracy of the measured data, may vary across the USA, adding another layer of complexity. It is reassuring from our study that exposures defined using the finer geographic scale (e.g., closer to true exposure) were similar to those defined by the coarser scales, though some of this similarity can be attributed to estimates calculated from the same monitors. Thus, it is not clear that the apparent benefit from modeling—greater geographic specificity—is necessarily worth increased uncertainty in the resulting data.

Neither the air monitoring data nor the NHIS are collected to study the effects of air quality on health outcomes. Air monitoring data is collected for regulatory purposes; the decisions that lead to the siting of monitors and their potential effects on these and other epidemiological results were not discussed here. The NHIS is designed to provide nationally representative information for a variety of health indicators. Care needs to be taken when combining these data sources for epidemiological health studies. In our evaluation of the combined data, we found differences in study samples by linkage method. For chronic outcome studies of the effects of fine particles, we did not find a compelling reason to limit the study samples to survey respondents within 5 miles of a monitor. Although not definitive, our results suggest that the possibly more precise exposure estimates may be outweighed by the potential bias and loss of sample size from restricting the sample in national studies.

Acknowledgments

Sources of financial support: Ms. Kravets was funded by an interagency agreement between NCHS/CDC and the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in the Department of Health and Human Services, USA, ASPE SP 05-039.

References

Basu, R., Woodruff, T.J., Parker, J.D., Saulnier, L., Schoendorf, K.C., 2004. Comparing exposure metrics in the relationship between $PM_{2.5}$ and birth weight in California. J. Exposure Anal. Environ. Epidemiol. 14, 391–396.

Bell, M.L., McDermott, A., Zeger, S.L., Samet, J.M., Dominici, F., 2004. Ozone and short term mortality in 95 US urban communities 1987–2000. J. Am. Med. Assoc. 292, 2372–2378.

Benjamins, M.R., Hummer, R.A., Eberstein, I.W., Nam, C.B., 2004. Self-reported health and adult mortality risk: an analysis of cause-specific mortality. Soc. Sci. Med. 59, 1297–1306.

Budtz-Jorgensen, E., Keiding, N., Grandjean, P., Weihe, P., White, R.F., 2003. Consequences of exposure measurement error for confounder

identification in environmental epidemiology. Stat. Med. 22, 3089–3100.
Chestnut, L.G., Schwartz, J., Savitz, D.A., Burchfiel, C.M., 1991. Pulmonary function and ambient particulate matter: epidemiological evidence from NHANES I. Arch. Environ. Health 46, 135–144.
Darrow, L.A., Woodruff, T.J., Parker, J.D., 2006. Maternal smoking as a confounder in studies of air pollution and infant mortality. Epidemiology (Research Letter) 17, 592–593.
Dey, A.N., Bloom, B., 2005. Summary Health Statistics for U.S. Children: National Health Interview Survey, 2003. National Center for Health Statistics. Vital Health Stat. 10 (223).
Dominici, F., Daniels, M., Zeger, S.L., Samet, J.M., 2002. Air pollution and mortality: estimating regional and national dose-response relationships. J. Am. Stat. Assoc. 97, 100–111.
Eberhardt, M.S., Pamuk, E.R., 2004. The importance of place of residence: examining health in rural and nonrural areas. Am. J. Public Health 94, 1682–1686.
Ellenberg, J.H., 1994. Selection bias in observational and experimental studies. Stat. Med. 13, 557–567.
Greenland, S., Gustafson, P., 2006. Accounting for independent nondifferential misclassification does not increase certainty that an observed association is in the correct direction. Am. J. Epidemiol. 164, 63–68.
Ingram, D.D., Franco, S., 2006. 2006 NCHS Urban–Rural Classification Scheme for Counties. Available from: 〈http://www.cdc.gov/nchs/r&d/rdc_urbanrural.htm〉.
Jerrett, M., Burnett, R.T., Ma, R., Pope III, A.C., Krewski, D., Newbold, K.B., Thurston, G., 2005. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology 16, 727–736.
Kravets, N., Hadden, W.C., 2007. The accuracy of address coding and the effects of coding errors. Health Place 13, 293–298.
Krieger, N., Chen, J.T., Waterman, P.D., Rehkopf, D.H., Subramanian, S.V., 2002. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. Am. J. Epidemiol. 156, 471–482.
Kunzli, N., Tager, I.B., 1997. The semi-individual study in air pollution epidemiology: a valid design as compared to ecologic studies. Environ. Health Perspect. 195, 1078–1083.
Le, H., Batterman, S., Dombrowski, K., Wahl, R., Wirth, J., Wasilevich, E., 2006. A comparison of multiple imputation and optimal estimation for missing and uncertain urban air toxics Data. Epidemiology (Abstract) 17, S242.
Lee, C., 2002. Environmental justice: building a unified vision of health and the environment. Environ. Health Perspect. 110 (Suppl. 2), 141–144.
Lethbridge-Cejku, M., Vickerie, J., 2005. Summary Health Statistics for U.S. Adults: National Health Interview Survey, 2003. National Center for Health Statistics. Vital Health Stat. 10 (225).
National Center for Health Statistics, 2007. Health, United States, 2007. Hyattsville, MD.
Ostro, B.D., 1983. The effects of air pollution on work loss and morbidity. J. Environ. Econ. Manage. 10, 371–382.
Ostro, B.D., 1987. Air pollution and morbidity revisited: a specification test. J. Environ. Econ. Manage. 14, 87–98.
Ostro, B.D., 1989. Estimating the risks of smoking, air pollution, and passive smoke on acute respiratory conditions. Risk Anal. 9, 189–196.
Ostro, B.D., 1990. Associations between morbidity and alternative measures of particulate matter. Risk Anal. 10, 421–427.
Ostro, B.D., Rothschild, S., 1989. Air pollution and acute respiratory morbidity: an observational study of multiple pollutants. Environ. Res. 50, 238–247.
Pope III, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. J. Air Waste Manage. Assoc. 56, 709–742.
Samet, J.M., Dominici, F., Curriero, F.C., Coursac, I., Zeger, S.L., 2000. Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994. New Eng. J. Med. 343, 1742–1749.
Schenker, N., Raghunathan, T.E., Chiu, P., Makuc, D.M., Zhang, G., Cohen, A.J., 2006. Multiple imputation of missing income data in the National Health Interview Survey. J. Am. Stat. Assoc. 101, 924–933.
Schille, J.S., Adams, P.F., Coviarty Nelson, Z., 2005. Summary health statistics for the U.S. population: National Health Interview Survey, 2003. National Center for Health Statistics. Vital Health Stat. 10 (224).
Schwartz, J., 1989. Lung function and chronic exposure to air pollution: a cross-sectional analysis of NHANES II. Environ. Res. 50, 309–321.
Schwartz, J., 2001. Air pollution and blood markers of cardiovascular risk. Environ. Health Perspect. 109 (Suppl. 3), 405–409.
Sheppard, L., Levy, D., Norris, G., Larson, T.V., Koenig, J.Q., 1999. Effects of ambient air pollution on nonelderly asthma hospital admissions in Seattle, Washington 1987–1994. Epidemiology 10, 23–30.
RTI International, 2006. SUDAAN, version 9.01 (computer software). Available from: 〈http://www.rti.org/sudaan/index.cfm〉.
U.S. Environmental Protection Agency, 2003. National Air Quality and Emissions Trends Report, 2003 Special Studies Edition. Research Triangle Park, NC. Available from: 〈http://www.epa.gov/air/airtrends/aqtrnd03/〉.
U.S. Environmental Protection Agency, 2006. Air Quality System (AQS) database (US EPA web page). Available from: 〈http://www.epa.gov/air/data/aqsdb.html〉.
Waller, L.A., Gotway, C., 2004. Applied Spatial Statistics for Public Health Data. Wiley, Hoboken, NJ, pp. 104–108.
Willis, A., Jerrett, M., Burnett, R.T., Krewski, D., 2003. The association between sulfate air pollution and mortality at the county scale: an exploration of the impact of scale on a long-term exposure study. J. Toxicol. Environ. Health A 66, 1605–1624.
Wong, D.W., Yuan, L., Perlin, S., 2004. Comparison of spatial interpolation methods for the estimation of air quality data. J. Exposure Anal. Environ. Epidemiol. 14, 404–415.
Woodruff, T.J., Grillo, J., Schoendorf, K.C., 1997. The relationship between selected causes of postneonatal infant mortality and particulate air pollution in the United States. Environ. Health Perspect. 105, 608–612.
Zeger, S.L., Thomas, D., Dominici, F., Samet, J.M., Schwartz, J., Dockery, D., Cohen, A., 2000. Exposure measurement error in time-series studies of air pollution: concepts and consequences. Environ. Health Perspect. 108, 419–426.
Zeka, A., Schwartz, J., 2004. Estimating the independent effects of multiple pollutants in the presence of measurement error: an application of a measurement-error-resistant technique. Environ. Health Perspect. 112, 1686–1690.