Relating air quality and environmental public health tracking data

V. Boothe\textsuperscript{1}, W. F. Dimmick \textsuperscript{2} & T. O. Talbot\textsuperscript{3}
\textsuperscript{1}Centers for Disease Control and Prevention
\textsuperscript{2}US Environmental Protection Agency
\textsuperscript{3}New York State Department of Health

Abstract

Initiated in February 2004, the Public Health Air Surveillance Evaluation (PHASE) Project is a multi-disciplinary collaboration between the Centers for Disease Control and Prevention (CDC), the U.S Environmental Protection Agency (EPA), and three Environmental Public Health Tracking Network (EPHTN) state agencies. The objective of this project is to develop, evaluate, and demonstrate the advantages and limitations of different methods of generating air quality characterization data that could be systematically and routinely available to link with public health surveillance data as part of the Environmental Public Health Tracking Network.

Keywords: ozone, particulate matter, asthma, myocardial infarctions, case cross-over, exposure assessment.

1 Introduction

In September of 2000, the Pew Environmental Health Commission issued a report that called on legislators to create a federally supported Nationwide Health Tracking Network of local, state and federal public health agencies that tracks trends of priority chronic diseases and relevant environmental factors in all 50 states. According to the report, effective environmental health tracking requires a coordinated approach that identifies hazards, evaluates exposures, and tracks the health of the population [1]. The term “Tracking” was defined by the Commission as synonymous with the concept of public health surveillance which is, “the ongoing systematic collection, analysis, and interpretation of
outcome-specific data for use in the planning, implementation, and evaluation of public health practice” [2].

In 2002, Congress funded the Centers for Disease Control and Prevention (CDC) to begin developing a nationwide Environmental Public Health Tracking Network (EPHTN). As part of the process, CDC prioritized existing environmental hazard data based on criteria including ongoing, systematic collection; finely resolved spatial and temporal variables; availability in most states; and effective quality assurance and control procedures. O₃ and PM₂.₅ ambient air quality monitoring data ranked highest.

Before the air monitoring and public health surveillance data can be utilized as part of the EPHTN, spatial and temporal compatibility issues and limitations must be addressed. Therefore, CDC, EPA, and three EPHTN state agencies (New York State Department of Health (NYSDOH), Maine Department of Health and Human Services (MDHHS) and the Wisconsin Department of Health and Family Services (WDHFS)), have collaborated as part of the Public Health Air Surveillance Evaluation (PHASE) Project to identify and evaluate data sources, tools, and methods that can be used to generate daily surrogate measures of O₃ and PM₂.₅ exposures and routinely and consistently assess the associated health impacts for all populations. Conducting these analyses is not expected to produce new scientific knowledge on the causal relationship between health and air pollution. However, the results will provide additional information to assist in identifying appropriate air characterization generation methods and in standardizing the data, methods and tools for the National EPHTN.

This evaluation includes several methods for generating surrogate O₃ and PM₂.₅ exposure measures including: (i) monitor proximity based assessments, (ii) statistical interpolation of ambient air monitoring data; (iii) the Community Multi-scale Air Quality (CMAQ) Model; and (iv) statistical combinations of data. A limited evaluation of the use of satellite data for the generation of PM₂.₅ surrogate exposure measures is also part of this project. The evaluation criteria for data generated by each method includes geographical and temporal coverage and resolution; costs to generate, access, analyze and store the data; and the spatial and temporal compatibility of the air quality data with the health data.

The role of the three states was to link air quality data with health data and to develop statistical models for examining the associations between health outcomes and O₃ and PM₂.₅ exposure levels. The tools developed will help ensure that the Environmental Public Health Tracking (EPHT) programs have the capability of assessing the relationships between these pollutants and adverse health events, on a routine and comparable basis.

This paper is organized into three main sections. First is a description of the air quality characterization methods and data sources used to generate the surrogate exposure measures. The second section contains an overview of NYSDOH PHASE Project objectives, a description of the public health importance and data sources for the health measures, and a description of the linkage and analysis methods. The final section is a discussion of the implications of the PHASE Project results and tools for EPHT and other environmental health programs.
2 EPA air quality measures: basis and initial analysis

Since the 1950’s, air quality has been measured systematically with the resulting estimates used to identify areas with unhealthy air. Ambient air quality levels are measured across the U.S. in a comprehensive monitoring network [3]. In addition to monitoring data, ambient air quality levels are estimated through various models based on emission estimates, meteorological and other data. Specifically, the emissions and meteorology data are fed into the model and run through various algorithms that simulate the physical and chemical processes in the atmosphere to provide estimated concentrations of the pollutants [4]. These models can be used to predict air quality in locations where no monitors exist or to simulate the effects of various changes in emission levels. In recent years, remote sensing techniques such as satellites have also begun to measure air quality levels. These levels can be related to air quality levels measured by ground-based monitors. Based on this relationship, it is possible to estimate air quality at locations where no monitor exists [5]. It should be noted, however, that there are some significant limitations to the currently available satellite data including the fact that data are only available under cloud-free conditions [6].

For the PHASE project, EPA’s primary role was to develop and evaluate data sets produced using alternative methods to generate air characterization data for public health tracking programs. The EPA team focused on methods that used readily available data that could be generated on an on-going systematic basis. In addition, EPA is exploring ways to improve the usefulness of O₃ and PM₂.₅ estimates for public health practitioners.

2.1 Development of O₃ and PM₂.₅ surrogate exposure measures

The O₃ and PM₂.₅ data sets developed within EPA can be divided into four types: monitored concentrations; interpolated monitored concentrations; modelled concentrations; and statistically combined monitored and modelled concentrations. The data covered areas encompassing New York, Maine and Wisconsin by day of the year in 2001. The PM₂.₅ estimates represent a 24-hour integrated sample while the ozone estimates represent the concentration averaged over the 8-hour maximum period in each day. Although some PM₂.₅ samples are taken every day, most are taken every third day. In contrast, ozone is monitored each day but, for most locations, only for days in the ozone season for the region.

2.1.1 Ambient air network monitoring data

The ambient air measurements for ozone and PM2.5 are gathered across the U.S. in a comprehensive population-based network by State and local agencies. The network is founded on data quality objectives, standardized methodologies, and requirements for data distribution. The data from the monitoring networks are used routinely to provide daily public health information on the Air Quality Index (AQI) to the public through the AIRNow program [7]. Ambient air measurements are considered the “true” measure of air quality, although there are known uncertainties in physical/chemical measurements [8]. Spatial gaps
exist in ambient monitoring data especially for rural areas. Government agencies site most air quality monitors in urban areas. The National Park Service does monitor most national parks, but many rural areas are not monitored [9].

2.1.2 Ambient monitoring data with statistical interpolation

Statistical interpolation of monitoring data improves spatial and temporal coverage by providing estimates of air quality in unmonitored locations and time periods. For each day and each pollutant, EPA used the interpolation method of ordinary kriging (i.e., kriging without the use of covariates) to produce estimates of air quality in unmonitored locations. Kriging is a form of statistical modelling that interpolates data from a known set of sample points to a continuous surface. According to published literature, “kriging” is the most common geostatistical technique used in the air pollution research field [10]. A major advantage of kriging over other interpolation methods is the production of both predicted values and their standard errors (kriging variance) at unsampled locations. These standard errors quantify the degree of uncertainty in spatial predictions at unsampled sites, providing valuable information on where the interpolation is less reliable [11].

EPA’s procedures for generating the kriged data sets used spatial covariance functions to describe the similarity between the pollution concentrations at two different locations as a function of the distance separating them. After determining the optimal covariance function, this method predicted O₃ and PM₂.₅ concentrations at the geographic centroid of the 4, 12, and 36 grid cell and the ZIP Code within each state. For PM₂.₅, time series interpolation was used to obtain predictions for the days on which no observations were available. A natural cubic spline was fitted to execute the interpolation.

2.1.3 Air quality modelling

The Models 3/Community Multiscale Air Quality (CMAQ) [12] was used to generate O₃ and PM₂.₅ concentrations at 36 km grid cells for the continental U. S. for all of 2001. The modelled estimates provide widespread spatial and high temporal resolution but have the potential for higher measurement error relative to monitoring data. The data provided to each state included daily maximum 1-hour and 8-hour O₃ along with daily average PM₂.₅ concentrations. In the spring of 2005, NOAA and EPA began generating routine CMAQ ozone data at 12km grid cells for Air Quality Index forecasts in the north-eastern portion of the U. S.

2.1.4 Statistical combination of these methods to generate a robust data set

EPA also produced O₃ and PM₂.₅ data sets by combining air monitoring and CMAQ data using hierarchical Bayesian methodologies. The method quantifies the uncertainties in both data sets to produce an improved estimate of error in the composite surrogate exposure estimates. It also draws on strengths of each data source while giving more weight to precise monitoring data in areas where monitoring exists, and relying on model output in non-monitored area.

2.1.5 Satellite data

Aerosol optical depth data (AOD), obtained from NASA for the Terra satellite for 2001, can be used to estimate PM₂.₅ daily concentrations for select days. The
algorithms and experimental uncertainties associated with using satellite data in the statistical combination of various datasets are still being investigated by EPA and NASA. However, satellite data, in combination with monitored and modelled data may be an important future EPHT data source. Satellites are unique in their ability to provide information regarding the origination and duration of exposure for large PM$_{2.5}$ transport events across the U.S. [13].

2.2 Comparison of selected output maps

The four types of data sets were provided in electronic form and graphical form to the PHASE Project state participants. The maps in Figure 1 illustrate each kind of data for one day. The top left map uses a color code to display the ozone ambient measurements for June 11. Each point represents a monitoring location. The top right map displays the CMAQ ozone predictions for June 11. The bottom left map displays he statistically interpolated ambient measurements for June 11 and the bottom right map showing the results of the Hierarchical Bayesian method, as applied to the ambient measurements and the CMAQ predictions. It is not prudent, at this time, to make comparison among the methods based on observations for a single day.

![Figure 1: Display of ozone levels based on four methods.](image)

2.3 Initial evaluation

In evaluating the relative merit of the four air characterization methods, there are a number of factors that need to be considered. Qualitatively, known inherent limitations to spatial and temporal coverage are being considered in evaluating the methods. Additionally, the ease/speed of implementation along with the costs in software, necessary computer resources and required expertise for each method are being assessed.

The best air quality characterization methods would be those that fill in missing O$_3$ and PM$_{2.5}$ data with the least variability and error and provide air
quality estimates that correlate closely with the ambient air monitored concentrations observed over time and space. The selection of a method would be based on the quantitative evaluations, on the qualitative evaluations where quantification is not possible, and the utility of the ambient estimates for public health assessments.

3 Evaluating relationships between health and air quality

The health departments of New York State, Wisconsin and Maine are focusing their efforts on understanding the performance of air characterization methods within the EPHT framework (linking health and environmental exposure data to drive public health action).

NYSDOH is working with the PHASE Team to examine how the relationships between O\textsubscript{3} and PM\textsubscript{2.5} and asthma exacerbations and acute cardiovascular effects vary based on choice of air characterization method and geographic scale. This component of the PHASE Project is designed to demonstrate the relative performance and potential public health utility of each of the air characterization generation methods. This analysis is ongoing and results will require careful interpretation. It is possible that there will be little or no difference in the analysis results between the alternative methods. This would indicate that correlation of the estimated concentrations with the monitored concentrations and the qualitative factors would be the most important criteria for method selection. On the other hand, if the analysis results are substantially different among the methods, then one or more of the methods may be clearly less accurate, and can thus be ruled out.

3.1 Geographic scale

The geographic scales of air quality and health data are important for several reasons. (1) If the health data is highly aggregated, a finer geographic resolution of air quality data may not provide any meaningful benefit and may not warrant the significant resources that might be needed to provide it. If, however, residential addresses are available and cases can be located to their street address, air quality data will likely be more useful at the finer scale. (2) PM\textsubscript{2.5} levels can vary considerably at a local level. However, there may not be as much benefit to finely resolved geographic data for O\textsubscript{3} because the concentrations are often relatively homogenous over large geographic regions.

3.2 Health data

To examine the relationship between the O\textsubscript{3}, PM\textsubscript{2.5} and adverse health effects, two health outcomes, acute myocardial infarction (AMI) and asthma exacerbations were selected. The choice was based on the ready availability of data in Maine, Wisconsin and New York, and scientific literature showing associations between these health conditions and ambient air pollution.

AMI is a leading illness and cause of death among adults. Epidemiologic studies have shown associations between its incidence and the ambient concentrations of airborne particulate matter, especially PM\textsubscript{2.5} [14, 15].
Asthma is a major illness and cause of disability; among children, it is the number one cause of emergency room visits, hospital admissions and doctor's office visits. During 1980 - 1996, the prevalence increased; since 1995, there has been an increase in outpatient visits and emergency department visits, and a decrease in mortality and hospitalizations [16], New York residents had an average of 42,725 asthma hospitalizations in 2001, and 358 deaths. Rates are associated with the levels of both particulate matter and ozone [17, 18]. Hospitalization data is being used for both AMI and asthma outcomes. For asthma, an expectation was that a large proportion of cases would be seen in emergency departments, without a resulting inpatient admission. However emergency room data were not available for the study period in New York. (Only the Maine Department of Health and Human Services was able to analyze emergency room data for this project.) For hospital admission data, NYSDOH is using the Statewide Planning and Research Cooperative System (SPARCS) [19]. Because the proportion of fatalities prior to arrival at a hospital may be significant for AMI, NYSDOH included AMI mortality data in the analyses. These data were obtained from the New York State Vital Records.

3.3 Linking health and air quality data

The health and air quality data were linked based on common geography. NYSDOH used automated geocoding software [20] to assign geographic coordinates based on the street address of the cases. The software matched 73 percent of the hospital discharge cases and 79 percent of the mortality cases based on street address. The ungeocoded cases were assigned the coordinates of the population-weighted centroids of their ZIP Codes. Each case was then assigned to a 4, 12 or 36 km grid cell based on the grid coordinates.

Table 1: Proportion of myocardial infarction cases assigned to the incorrect grid cell by method of grid cell assignment.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Km</td>
<td>49.5</td>
<td>41.5</td>
<td>38.7</td>
</tr>
<tr>
<td>12 Km</td>
<td>20.1</td>
<td>16.1</td>
<td>15.4</td>
</tr>
<tr>
<td>36 Km</td>
<td>7.3</td>
<td>5.4</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Different methods of assigning geographic coordinates based on ZIP Codes were assessed. Using population-weighted ZIP Code centroids led to some cases being assigned to incorrect grid cells. The population weighted centroid approach was found to mis-assign 41.5 percent cases at the 4 km scale, 16.1 percent at the 12 km scale and 5.4 percent at the 36 km scale. However, the percentage mis-assigned would have been greater if the geographic centroid of the ZIP Code was used rather than the population weighted centroid (Table 1).
These geographic centroids are easier for the user to assign since they are provided by most GIS systems with no need for software programming by the user. When cases were assigned to the grid cell where the largest proportion of a ZIP Code population was located, the results improved slightly (Table 1). However, this method is more complex to implement, while providing little reduction in the error rate. Because ZIP Code boundaries do change with time, assigning case coordinates based on ZIP Codes will lead to problems in tracking the air pollutant data over time. Also, if air quality areas must align with current ZIP Code boundaries, they would have to change continually.

### 3.4 Statistical analysis of health and air quality

The association between the health outcomes and air quality levels are being examined using a case-crossover design. The case-crossover design [21] compares air quality levels just prior to a “case” day on which a person is hospitalized with the levels on one or more “control” days when the person was not hospitalized. The method is well suited for estimating the association between short-term exposures and acute health events shortly thereafter, such as asthma attacks and AMI. Since each case provides its own control, behavioral and socioeconomic risk factors as well as unknown personal risk factors are controlled for by design.

The latest research [22] recommends the time-stratified case-crossover design. In this strategy, time is divided into non-overlapping “windows” or referent periods (a referent period of 28 days will produce 13 “windows” in one calendar year). Cases and the air quality levels just before case days are compared to air quality levels just before control days within the same “window”. Within a window, controls are commonly selected on the same day of the week as cases, to control for the fact that healthcare utilization, personal activities, and concentrations of air pollution vary by day of the week. The challenging part of this design is the selection of the best window size. Optimally, the control periods are far enough from the case period that any health effects do not carry over between the two, but not so far apart that seasonal and sociodemographic factors differ. Their judicious selection will often control for time-varying confounders. To gain insight into whether the effect of air pollution was distributed over several days, air pollution on the day of the hospitalization, as well as one, two and three day lags before hospitalization, and averages over as many as three days, are included in the analyses.

### 3.5 Evaluation and comparison of analysis results

The strength of association and the uncertainty in the estimates from the case-crossover analyses will be compared in tables, charts and graphs to assess how the results differ among air characterization methods. To assess the potential for misclassification due to aggregation, the results will be compared according to level of geographic resolution. An evaluation of whether the distribution of the effect over time, that is to say the lag structure, is consistent among the air characterization methods will also be considered.
4 PHASE products - implications for EPHT and research

Ideally, a national EPHT Network would facilitate nationwide sharing of standards-based health and environmental data. However, it will take time to address confidentiality issues, negotiate data sharing agreements, and to institute secured access and requisite security procedures. In the interim, the three PHASE states are using standardized case definitions, EPA generated surrogate air quality exposure measures, and similar approaches for data linkage and case-crossover analysis. The expectation is that this effort will further EPHT goals by making it possible for states to directly compare their analytic results.

A guide containing case definitions; descriptions and limitations of the health and air quality data; and the linkage and analysis methods used in the PHASE Project will also be made available to public health professionals and researchers. Comparison of the results generated through widespread utilization of guide would greatly expand the current scientific body of knowledge regarding the link between exposure to air pollution and health.

Acknowledgement

This report does not constitute an endorsement of authors, or organizations by CDC. The views and opinions of these authors and organizations are not necessarily those of CDC or the Department of Health and Human Services (HHS).

References

[7] EPA AIRNow Website See (http://cfpub.epa.gov/airnow/)


