HEI

STATEMENT

Synopsis of Research Report 183, Parts 1 & 2

HEALTH EFFECTS INSTITUTE

New Statistical Methods for Analyzing Multiple Pollutants, Sources, and Health Outcomes

BACKGROUND

The National Research Council recommended in 2004 that the U.S. Environmental Protection Agency take steps to address the presence of a complex, multipollutant atmosphere in the process for reviewing and setting the National Ambient Air Quality Standards, which are currently based on single pollutants. One of those steps included improving statistical methods to evaluate how simultaneous exposure to multiple ambient air pollutants affects human health. Conventional statistical methods are not well suited to deal with high correlations among pollutants, differences in the composition of pollutant mixtures over time and space, or differences in how accurately a person's actual exposures to individual pollutant concentrations have been estimated. These factors can lead to errors in the estimation of the health effects associated with individual or multiple pollutants and the emission sources with which they may be associated.

In response to these concerns, the Health Effects Institute issued request for applications 09-1, "Methods to Investigate the Effects of Multiple Air Pollution Constituents," to fund development of innovative statistical methods that could be applied to real-world exposures and health problems. HEI funded three studies: the two studies led by Dr. Brent Coull and Dr. Eun Sug Park are described in the current report, and a third study led by Dr. John Molitor is expected to be published in 2016. Both Coull and Park proposed the use of Bayesian statistical methods, which essentially allow for the integration of prior knowledge or data about a problem with new data in the same analysis, thus allowing for a more comprehensive evaluation of available information, including the characterization of uncertainty in the analytic process. Both investigative teams explored joint modeling of exposure and health outcomes in contrast to the conventional two-stage approach in which exposures are estimated first and then input to the health effects analysis. Each team developed and demonstrated their methods using data on health outcomes related to short-term changes in particulate matter composition and levels.

What These Studies Add

- Both the Coull and Park studies have advanced the use of Bayesian statistical methods to address shortcomings in the ability of existing approaches to disentangle the roles of individual pollutants in short-term studies of complex multipollutant exposures and their health effects.
- Coull and associates developed methods to identify which key pollutants within a simple mixture are most closely associated with adverse health outcomes, to accommodate a variety of exposure–response relationships, and to characterize uncertainty in the estimated health effects more fully.
- Park and colleagues extended existing methods for characterizing relationships between emission sources and health by (1) allowing for the contributions from sources to be correlated and (2) making sure that the health effects estimates account for various uncertainties in estimating the source contributions. The team also developed enhanced models that could take into account correlations among pollutants from more than one monitoring location and estimate source contributions at locations of interest for which monitoring data may be lacking.

This Statement, prepared by the Health Effects Institute, summarizes a research project funded by HEI and conducted by Dr. Brent A. Coull of the Harvard T.H. Chan School of Public Health, Boston, MA, and by Eun Sug Park of the Texas A&M Transportation Institute, College Station, TX, and their colleagues. The complete report, *Development of Statistical Methods for Multipollutant Research* (© 2015 Health Effects Institute), can be obtained from HEI or our Web site (see last page).

STUDY BY COULL AND COLLEAGUES: STATISTICAL LEARNING METHODS FOR THE EFFECTS OF MULTIPLE AIR POLLUTION CONSTITUENTS

Approach

Coull and his colleagues developed methods that simultaneously select pollutants to include in the models, provide flexible approaches to estimating exposure–response relationships (for example, allowing them to be nonlinear), identify interactions among pollutants (for example, additive or synergistic effects), and allow for the quantification of uncertainty (by using Bayesian kernel machine regression [BKMR] methods). These methods involve a joint-estimation approach in which the identification of important exposure variables is, in a sense, "supervised," or influenced by, the data on health outcomes.

The investigators formally developed and tested different features of their BKMR methods first in three simulation studies and then in two real-world health and exposure data sets. The simulation studies were designed to compare the performance of their methods with those of more conventional ones in a range of plausible scenarios, defined by the investigators, involving different numbers of important pollutants or sources, nonlinear as well as linear exposure–response relationships, and different kinds of interactions among the exposure constituents. An important feature of their simulations was that their air pollution data sets were generated from actual $PM_{2.5}$ constituent data measured at a Boston monitoring site, thereby retaining the realistic joint distributions and correlations among the multiple pollutants. They then applied their methods to data from two previously published Boston studies — an epidemiologic study that had evaluated changes in blood pressure after short-term exposure to constituents of $PM_{2.5}$ in patients 70 years of age and older, and a toxicologic study with laboratory dogs. These studies also relied on pollutant data from the same Boston monitoring site.

Results and Discussion

In its independent review of the study, the HEI Health Review Committee noted that the statistical approach developed by Coull and associates had carefully addressed a number of the challenges researchers face in dealing with multiple pollutants and sources when using more conventional statistical methods. Their methods also allowed the uncertainties associated with statistical modeling to be more fully reflected in the health effects estimates, providing useful insight into the degree of confidence in those estimates. The Committee thought the investigators had provided a strong theoretical basis for their approach and that their simulations and real-world applications were well chosen to demonstrate the practical use of their methods. A strength of those choices was that the simulated pollutant data sets were generated from the same pollutant data used in the two real-world studies and thus made the simulation results more relevant for comparisons with those of the previous epidemiologic and toxicologic analyses.

The methods worked as expected in the simulations but with some limitations in the analyses of the epidemiologic and toxicologic data sets. In the simulated data sets, the methods characterized exposure–response relationships in various forms and were more likely to correctly identify the pollutants used to predict the adverse health outcomes than more standard methods. However, as with conventional statistical methods, it remained challenging to identify correctly the relative importance of an individual pollutant's contribution to health outcomes when high degrees of correlation existed among the suite of pollutants. A limitation in the data sets for older adults and for dogs was that they did not have either the size or complexity to represent the kinds of interactions among pollutants or the nonlinearities in the exposure–response relationships that would be necessary to test those features of the methods.

The Committee thought it was likely that the methods developed by Coull and colleagues were more systematic and transparent in identifying the absence of interactions or nonlinearities than conventional data analysis approaches would be. In conventional analyses, investigators would ordinarily need to cycle through a series of models to test whether interactions were present, a process that would require a number of analytic choices and raise issues of multiple testing and possibly false-positive findings. Methods such as the ones developed by Coull and associates could be helpful in minimizing the ad hoc nature of this process. However, as is the case in the development of any statistical approach, these methods need to be applied in a broader range of scenarios before their usefulness in real-world practice can be ascertained.

STUDY BY PARK AND COLLEAGUES: DEVELOPMENT OF ENHANCED STATISTICAL METHODS FOR ASSESSING HEALTH EFFECTS ASSOCIATED WITH AN UNKNOWN NUMBER OF MAJOR SOURCES OF MULTIPLE AIR POLLUTANTS

Approach

Park and her colleagues developed a set of methods to analyze daily variations in health and source-apportioned air pollution (time-series data). In their first specific aim, Park and associates developed a Bayesian modeling approach that estimated the number of sources and the contributions of each to exposures at the same time as it estimated the effects of those exposures on human health outcomes. Their joint modeling approach incorporated uncertainties in the source-apportionment process into the final estimates of uncertainty in the health effects estimates. More specifically, their analysis allowed them to examine the impact of correlations among source contributions to exposure as well as incorporating uncertainty from other modeling assumptions into their final estimates of health effects. In standard applications of source apportionment, the number of sources and how much each one contributes to exposure is assumed to be known without error; if this assumption is not true, it could lead to misspecification of how health effects are attributed to different categories of sources.

In their second specific aim, the team developed methods for modeling the contributions of multiple sources to exposures that could handle more complex data and model structures than conventional source-apportionment methods. That is, they designed methods to incorporate data from more than one monitoring location to account for spatial correlations among multiple pollutant measurements collected at several locations, and to estimate source contributions at locations where no data were available.

For each of the specific aims, the investigators evaluated their methods in two steps. First, they conducted simulation studies in which the characteristics of the data, sources, and health effects were specified by the investigators (that is, they did not draw directly from actual collected data as did Coull and associates). Second, they applied their methods to real-world data sets in order to gain a more practical perspective. To test their methods for estimating source-related health effects (Aim 1), the investigators studied the associations of daily $PM_{2.5}$ speciation data with respiratory mortality in Houston, Texas, and with cardiovascular mortality in Phoenix, Arizona. In both of these cases, $PM_{2.5}$ data had been collected at individual monitoring sites. To test their more complex source-apportionment models (Aim 2), they examined data on volatile organic compounds from nine monitoring sites in Harris County, Texas, near Houston.

Results and Discussion

In its independent review of the study, the HEI Health Review Committee concluded that Park and colleagues had tackled an extremely challenging technical problem and, in spite of its difficulty, conducted a high-quality study that has provided a meaningful extension of existing source-apportionment approaches. Useful innovations include the joint estimation of sources and health effects, while accounting for uncertainty in source-apportionment models, allowing for spatial correlations among data from multiple monitoring locations, and estimating source contributions to exposure at locations without monitoring data. However, implementing the joint models and the spatial multivariate receptor models is challenging because they require data to be in a specific form that is often not available in existing data sets.

The Committee thought that Park and associates had raised important scientific issues with existing methods and developed new approaches to address them. The investigators properly developed and tested their methods in both simulations and applications to real-world data sets. The simulations performed well under the range of conditions evaluated. The applications of the methods to real-world data also provided evidence that the methods appear to work as intended in identifying sources and source-related health effects, albeit with differences from other published studies that need further investigation. Uncertainties in the health effects estimates tended to be larger, which reflected the more comprehensive accounting for uncertainty in the work. The Committee thought the enhanced modeling methods developed as part of the investigators' second aim appeared to be a useful innovation and were able to predict source contributions at unmonitored locations. However, for any of these methods to gain widespread scientific applicability, the Committee advised that they need to be applied in other settings, particularly ones that would allow comparisons with other studies that use more conventional approaches.

CONCLUSIONS

The HEI Health Review Committee concluded that each of the studies by Coull and Park and their colleagues addressed important but separate questions in multipollutant research. Both investigator teams followed logical steps in developing their methods from the conceptual underpinnings and then applying the methods to simulated and real-world data sets. Each team made considerable progress in demonstrating the feasibility and applicability of their approaches.

Challenges still remain, however, and further work is necessary to apply and evaluate the proposed methods. Although both sets of methods are already quite computationally demanding, they need to be evaluated in a broader range of real-world settings representing different levels of data complexity. They have also not yet been evaluated in studies of long-term exposure to air pollution. Where possible, these evaluations should include side-by-side comparisons of the new approaches against the more conventional two-stage approach. Such direct comparisons could help to determine whether the additional complexity of these new methods will lead to better understanding of how pollutant mixtures and their sources may contribute to effects on human health and, ultimately, to better decisions about how to control them.

Table of Contents

Development of Statistical Methods for Multipollutant Research

HEALTH EFFECTS INSTITUTE

101 Federal Street, Suite 500
Boston, MA 02110, USA
+1-617-488-2300 phone
+1-617-488-2335 fax

pubs@healtheffects.org
www.healtheffects.org