HEALTH EFFECTS INSTITUTE

STATEMENT

Synopsis of Research Report 152

Evaluating Heterogeneity in Indoor and Outdoor Air Pollution Using Land-Use Regression and Constrained Factor Analysis

BACKGROUND

Epidemiologic studies of exposure to air pollution have typically relied on data from centrally located ambient air quality monitors. However, such data are not sufficient for capturing the spatial variability of pollutant concentrations at the local scale, in particular at the within-city, or intra-urban, scale at which traffic-related air pollution is both highest and most variable. The ideal approach would be to measure each individual's personal exposure to traffic-specific pollutants over time, but this is difficult, intrusive, expensive, and generally not feasible for very large populations. Investigators have consequently sought ways to predict, or to model, individual-level exposures from more readily available data.

Numerous studies have had to rely on a variety of surrogates for such exposures: measured levels of individual pollutants previously associated with traffic emissions (e.g., carbon monoxide, nitrogen dioxide [NO_2], fine particulate matter with an aerodynamic diameter ≤ 2.5 µm [$PM_{2.5}$], benzene, and elemental carbon [EC]) and various measures of traffic density or of proximity to traffic. More complex techniques, such as land-use regression (LUR) models, have been increasingly developed to take advantage of data available from geographic information systems (GIS) — nearby land-use patterns, traffic, physical site characteristics, housing, and other variables — that have been hypothesized to provide additional information useful for predicting concentrations of traffic-related pollutants. Each type of surrogate has limitations in predicting levels of personal exposures to traffic-related pollutants. HEI Special Report 17, a critical review of the literature on traffic-related air pollution, found, in particular, that many of the simpler surrogates do not perform well. The resulting error in individuals' exposures can in turn affect the size and significance of health outcome findings from observational epidemiologic studies.

Dr. Jonathan I. Levy of the Harvard School of Public Health was awarded funding from HEI under Request for Applications 04-5, the Walter A. Rosenblith New Investigator Award. In his application, "Using Geographic Information Systems (GIS) to Evaluate Heterogeneity in Indoor and Outdoor Concentrations of Particle Constituents," Levy had proposed an approach to extend and improve upon existing GIS-based methods for predicting intra-urban exposures. An underlying goal of his study was therefore to explore ways to reduce exposure-measurement error and to improve the accuracy and precision of the associations reported in epidemiologic studies of air pollution.

APPROACH

Levy and colleagues conducted a study linked to a prospective birth cohort study of factors that might contribute to the development of asthma, the Asthma Coalition for Community, Environment, and Social Stress study in Boston, Massachusetts. Among the several factors under investigation were indoor and outdoor exposures to air pollutants, including those potentially related to traffic. Levy and colleagues collected detailed air quality measurements at a set of homes selected to reflect a range of potential exposures to traffic and of neighborhoods broadly representative of Boston. From 2003 through 2005, the investigators collected short-term NO_2 and $PM_{2.5}$ samples simultaneously indoors and outdoors at each home during two seasons. The particle filters used to collect the samples were analyzed for EC and for individual elements using two different analytical methods. The investigators also obtained hourly NO_2, $PM_{2.5}$, EC, and meteorological data for the study period from centrally located Massachusetts Department of Environmental Protection monitors, to provide data on the variation in background pollutant levels over time.

This Statement, prepared by the Health Effects Institute, summarizes a research project funded by HEI and conducted by Dr. Jonathan I. Levy at the Harvard School of Public Health, Boston, Massachusetts, and colleagues. The complete report, *Evaluating Heterogeneity in Indoor and Outdoor Air Pollution Using Land-Use Regression and Constrained Factor Analysis* (© 2010 Health Effects Institute), can be obtained from HEI or our Web site (see last page).

The investigators collected several additional types of data to support development of their LUR models. They utilized existing GIS data on road networks, traffic counts, and population density to characterize proximity to and potential density of traffic in the vicinity of each home. They obtained additional data on local land use and on the age of each home, its living area, building materials, heating system, and whether or not it had air conditioning. Investigators administered a standardized questionnaire to participants at each home to obtain data on occupant behaviors and home characteristics that have been shown previously to indicate indoor sources of pollutants or to influence ventilation in the home.

Levy and his colleagues then undertook a series of systematic analytic approaches to predicting concentrations of $PM_{2.5}$, EC, and NO_2 measured at each of the homes in the study and to understanding their potential sources. Using multiple variables drawn from their indoor and outdoor residential monitoring data, GIS-based land-use data, and questionnaire data, they first developed separate GIS-based LUR models to predict concentrations of $PM_{2.5}$, EC, and NO_2 measured outdoors and indoors at the residences. Second, using constrained factor analysis, a source apportionment technique, they analyzed the particle components and NO_2 measurements to identify potential source categories for pollutants measured outdoors and indoors at the homes in the study. The third step in their approach was to apply LUR analysis to the results of their source apportionment analyses; that is, they developed additional LUR models designed specifically to help explain variability in the source categories identified from their source apportionment analysis. By evaluating the extent to which they could successfully predict sources using particular GIS, land-use, and questionnaire data collected for the study, the investigators sought to corroborate their initial interpretations of the source apportionment analysis.

Finally, using simulation techniques, they conducted an analysis to assess how a variety of possible surrogates for indoor exposures, representing different levels of exposure-measurement error, could influence epidemiologic estimates of the relationship between indoor pollutant concentrations and reports of wheeze (a possible indicator of asthma) in a child's first year of life. They compared the performance of their indoor LUR models for $PM_{2.5}$, NO_2, and EC to that of surrogates based on single variables that had performed well in their model development process ("good exposure surrogates") and to that of surrogates based on traffic indicators that had not performed well in their analysis, but that had been used in studies reported by other investigators ("poor exposure surrogates"). They ran their simulations using three scenarios for the strength of the "true" associations between individual exposure and wheeze.

RESULTS

Levy and his colleagues reported that their final multivariate outdoor LUR models performed reasonably well; the models were able to explain most of the variability in outdoor residential concentrations of EC, NO_2, and $PM_{2.5}$ (52%, 56%, and 76%, respectively). EC and NO_2 had stronger relationships with indicators for local traffic than did $PM_{2.5}$. The variation in pollutant levels over time, represented by measurements at the central site monitor and seasonal terms, explained more of the variability in $PM_{2.5}$ (68%) than in EC (30%) or NO_2 (33%), a finding consistent with other studies.

They reported less success in the ability of their LUR regression models to predict indoor concentrations of the three pollutants. The indoor LUR models could explain only 20%, 21%, and 36% of the variation in indoor NO_2, EC, and $PM_{2.5}$ levels, respectively. They found identifying traffic terms with strong explanatory power to include in their models to be particularly challenging. When ventilation terms were introduced into the models, the explanatory power of the models increased slightly, to 25%, 32%, and 40%, respectively. The investigators reported that the ratios between the indoor and outdoor concentrations of individual particle constituents varied substantially among the different constituents, which later enabled some distinctions to be made between their potential indoor and outdoor sources.

From their source apportionment analysis using outdoor pollutant concentrations, Levy and his colleagues indentified five broad source categories: long-range transport; brake wear and local traffic; diesel exhaust; fuel oil combustion; and road dust and resuspension. Their analysis of measured indoor pollutant concentrations suggested six possible source categories, three interpreted to have origins outside the home — long-range transport, fuel oil/diesel combustion, and road dust and resuspension — and three interpreted to have indoor origins — indoor combustion, indoor smoking, and indoor cleaning.

Levy and colleagues had mixed success in their efforts to use LUR models to explore more fully the potential predictors for the source categories they had identified. In general, the LUR models they developed to predict the outdoor source categories had weaker explanatory power than those they had developed earlier to predict the levels of individual pollutants. The one exception was the investigators' LUR model for long-range transport, which was able to explain 69% of the variation observed. The strong performance of this model was consistent with that of the earlier LUR model for outdoor $PM_{2.5}$, since long-range transport was most closely associated with $PM_{2.5}$ measured at the central site monitor. The LUR model for predicting long-range transport indoors also performed the best. Most of the variation (68%) in the source category was explained by a term representing a combination of $PM_{2.5}$ data from the central site monitor and a variable obtained from the questionnaires that was indicative of greater ventilation in the homes (i.e., open windows). The LUR models for the remaining indoor source factors had very little explanatory power, in most cases substantially less than the LUR models for outdoor source factors.

From their simulation analysis, Levy and colleagues reported that the risks of wheeze estimated using exposures to individual pollutants based on their indoor LUR models were closer to the "true" risks than those estimated using either the simpler "good" or "poor" surrogates for exposure. That is, there was less bias and less uncertainty in the predicted risk estimates relative to the known risks used in the simulation. The models for NO_2 and $PM_{2.5}$ performed better than the one for EC. The investigators inferred from this simulation analysis that their LUR models predicted individual exposure levels with less exposure-measurement error than the individual surrogate approaches, and thus enhanced the power of the simulated epidemiologic study to detect the underlying association between wheeze and the pollutants.

CONCLUSIONS

Levy and his colleagues took advantage of a small but rich data source related to a study of childhood asthma in a major U.S. city to explore important exposure questions that are of broad interest to environmental health science. They undertook a number of challenging methodologic approaches to improving predictions of personal exposure to pollution from indoor and outdoor sources and thus to improving epidemiologic estimates of the effects of traffic-related air pollution on health. Their report marks one of the first efforts to combine LUR models with source apportionment analysis to characterize potential exposures to both indoor and outdoor sources. The HEI Health Review Committee praised the evident care and competence demonstrated by the investigators in their work.

The investigators' LUR analyses of outdoor pollutant levels performed reasonably well, explaining most of the variation in concentrations of $PM_{2.5}$ and to a lesser extent in NO_2 and EC. Their results were consistent with previously published findings showing that broader-scale temporal variation, represented by measurements at the central site monitor, is an important determinant of local $PM_{2.5}$ levels. Spatially distributed factors, such as traffic, population density, and other land-use covariates, were more influential in predicting EC and NO_2 variation in the models.

Development of LUR models to predict indoor concentrations, as a closer proxy for personal exposure, proved to be a much greater challenge. The predictive value of the indoor LUR models was generally poor; however, the authors' exploration of the possible explanations and implications of this finding is thorough and informative. Their findings that the indoor LUR models' performance was poorer when important indoor sources were present is noteworthy, as is the finding that the performance of an indoor LUR model can be improved by the relatively straightforward addition of a proxy term for ventilation taken from questionnaire data (i.e., "open windows").

An ultimate, and the most innovative, goal of the study was to see if LUR modeling and source apportionment analysis together would provide additional insight about the sources contributing to outdoor and indoor concentrations of pollutants and help explain why their contributions might differ at individual homes. The analyses did provide some confirmation for the investigators' major source interpretations (long-range transport, traffic, fuel oil combustion). However, they were most successful at explaining variation in sources that are already reasonably well understood. In particular, both the outdoor and indoor LUR models performed best at predicting variability in the source identified most closely with $PM_{2.5}$ concentrations at the central site monitor, long-range transport. As for the indoor LUR models developed to predict individual pollutant concentrations, incorporating a proxy for ventilation improved the performance of the indoor LUR models designed to predict variation in this source.

The ultimate challenge for studies of this nature is to provide some demonstration that the increased sophistication of the modeling provides a sufficient improvement over simpler approaches to warrant the additional data and computational requirements it imposes. The investigators' simulation analyses, in which they explore the implications for the power of epidemiologic studies of using different surrogate measures of individual exposure, are a useful step in that direction. Their conclusion that even the relatively poor estimates of exposure provided by the LUR models might reduce measurement error and thus improve effects estimates in future studies warrants further scrutiny. Even when a poor surrogate is outperformed by a prediction model, the surrogate may be the epidemiologist's "best buy" if the extent of improved performance is outweighed by the costs of collecting the data necessary for the prediction model.

Research Report 152

Evaluating Heterogeneity in Indoor and Outdoor Air Pollution Using Land-Use Regression and Constrained Factor Analysis

Jonathan I. Levy, Jane E. Clougherty, Lisa K. Baxter, E. Andres Houseman,
and Christopher J. Paciorek

INVESTIGATORS' REPORT

CRITIQUE by the Health Review Committee

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

Recycled Paper