OCTOBER 2006

HEI Communication 12

HEALTH EFFECTS INSTITUTE

Internet-Based Health and Air Pollution Surveillance System

Scott L Zeger, Aidan McDermott, Francesca Dominici, Roger Peng, and Jonathan Samet

HEALTH
EFFECTS
INSTITUTE

The Health Effects Institute is a nonprofit corporation chartered in 1980 as an independent research organization to provide high-quality, impartial, and relevant science on the effects of air pollution on health. To accomplish its mission, the Institute

- Identifies the highest-priority areas for health effects research;
- Funds and oversees the conduct of research projects;
- Provides intensive independent review of HEI-supported studies and related research;
- Integrates HEI's research results with those of other institutions into broader evaluations; and
- Communicates the results of HEI research and analyses to public and private decision makers.

Typically, HEI receives half of its core funds from the US Environmental Protection Agency and half from the worldwide motor vehicle industry. Frequently, other public and private organizations in the United States and around the world also support major projects or certain research programs. HEI has funded more than 250 studies in North America, Europe, and Asia that have produced important research to inform decisions regarding carbon monoxide, air toxics, nitrogen oxides, diesel exhaust, ozone, particulate matter, and other pollutants. The results of these studies have been published in more than 200 Research and Special Reports.

HEI's independent Board of Directors consists of leaders in science and policy who are committed to the public–private partnership that is central to the organization. The Health Research Committee solicits input from HEI sponsors and other stakeholders and works with scientific staff to develop the Five-Year Strategic Plan, select research projects for funding, and oversee their conduct. The Health Review Committee, which has no role in selecting or overseeing studies, works with staff to evaluate and interpret the results of funded studies and related research.

All project results and HEI Commentaries are widely communicated through HEI's website (*www.healtheffects.org*), annual conferences, publications, and presentations to legislative bodies and public agencies.

HEALTH EFFECTS INSTITUTE

CONTENTS

Communication 12

Internet-Based Health and Air Pollution Surveillance System

Scott L Zeger, Aidan McDermott, Francesca Dominici, Roger Peng, and Jonathan Samet

Bloomberg School of Public Health, Johns Hopkins University, Baltimore MD

PREFACE

CONTRIBUTORS

PROJECT REPORT

COMMENTS

Publishing history: This document was posted as a preprint on *www.healtheffects.org* on October 20, 2006, and finalized for print in December 2006.

Citation for whole document:

Zeger SL, McDermott A, Dominici F, Peng R, Samet J. 2006. Internet-Based Health and Air Pollution Surveillance System. Communication 12. Health Effects Institute, Boston MA.

When specifying a section of this report, cite it as a chapter of the whole document.

PREFACE

HEI Communication 12, *Internet-Based Health and Air Pollution Surveillance System*, SL Zeger et al

HEI Communication 12 describes a project by Dr Scott Zeger and colleagues of the Johns Hopkins Bloomberg School of Public Health that was funded by HEI with the purpose of making data and software from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS*) available to a wide range of interested parties. This Communication contains the investigators' Project Report, which describes the Internet-Based Health and Air Pollution Surveillance System (iHAPSS), followed by Comments from some members of the HEI Health Research and Review Committees and other experts.

HEI has always recognized that to be credible, the science it funds and the data underlying it must be transparent to other interested parties. In response to increased interest in its science, in particular the Reanalysis of the Harvard Six Cities and American Cancer Society Studies (Krewski et al 2000), in the late 1990s HEI adopted its first policy to ensure public access to data from studies it funds. Around that time, public access became a higher priority for the broader scientific community because of (1) the National Academy of Sciences recommendations on data issues in the natural sciences (National Research Council 1997) and (2) the Shelby Amendment (a provision inserted in the fiscal year 1999 Omnibus Appropriations Bill to allow broader access to federally funded research data). After the federal regulations for implementing the Shelby Amendment were issued, HEI formalized its Data Access Policy (Comments Appendix A) to ensure consistency with provisions in the Shelby Amendment. Beyond that, and in the spirit of transparency and fostering data access, HEI sought to identify more proactive ways to make data available to the public, especially for high-impact projects such as NMMAPS.

Increasing access to data enhances credibility by providing transparency about the research process and data interpretation. It also provides an opportunity for other researchers to validate and reanalyze the results, and to perform additional analyses. HEI has extensive experience conducting reanalyses of studies that have regulatory importance: for example, the Particle Epidemiology Evaluation Project (Samet et al 1995, 1997), the studies of railroad workers that evaluated their exposure to diesel emissions and incidence of lung cancer (Health Effects Institute 1999), the Reanalysis of the Harvard Six Cities and American Cancer Society Studies (Krewski et al 2000), and most recently the Revised Analyses of Time-Series Studies of Air Pollution and Health (Health Effects Institute 2003).

Because of the importance of NMMAPS in science and regulatory decisions, research groups interested in conducting additional data analyses have requested access to NMMAPS data. In response, Dr Zeger and colleagues submitted a proposal to create a website that would provide electronic access to NMMAPS data and to the software used for the data analyses. This approach was intended to facilitate study replication and new analyses. It would also reduce the burden on the NMMAPS investigators to respond to numerous individual requests for information. This model could potentially be applied to other research expected to yield results of high interest for regulatory and public health purposes.

As the project neared completion, the investigators and HEI assembled a group of sponsors and others from the community of expected users to assess the practical operation and ultimate utility of the website. After the project was completed, HEI approached this group and others in government, industry, and academia to provide written comments evaluating the website. Some members of both the HEI Health Research and Review Committees also provided written comments about the project.

Health Effects Institute

* A list of abbreviations and other terms appears at the end of the Project Report.

CONTRIBUTORS
Communication 12

HEALTH EFFECTS INSTITUTE

HEI Project Staff

Aaron Cohen *Principal Scientist, Project Oversight*

Annemoon van Erp, *Senior Scientist, Review Oversight*

Terésa Fasulo, *Science Administration Manager*

Robert O'Keefe, *Vice President*

Jane Warren, *Director of Science*

HEI Publications

Virgi Hepner, *Senior Science Editor*

Carol Moyer, *Consulting Science Editor*

Kasey Oliver, *Administrative Assistant*

Ruth Shaw, *Consulting Designer and Compositor, Cameographics Publications*

Project Evaluators

Ross Anderson, *Division of Community Health Sciences, St George's, University of London, and HEI Review Committee*

Ben Armstrong, *Public and Environmental Health Research Unit, London School of Hygiene and Tropical Medicine, and HEI Review Committee*

Barbara Glenn, *National Center for Environmental Research, US Environmental Protection Agency*

Bryan Hubbell, *Office of Air Quality Planning and Standards, US Environmental Protection Agency*

Dennis Kahlbaum, *Air Improvement Resource, Inc*

Timothy Ramsay, *McLaughlin Centre for Population Health Risk Assessment, University of Ottawa*

Paul Rathouz, *Department of Health Studies, University of Chicago*

Nancy Reid, *Department of Statistics, University of Toronto, and HEI Review Committee*

Howard Rockette, *Department of Biostatistics, University of Pittsburgh, and HEI Research Committee*

Margaret Round, *Environmental Health Assessment, Massachusetts Department of Public Health*

Ira Tager, *Division of Epidemiology, School of Public Health, University of California, Berkeley, and HEI Research Committee*

Heather Walton, *Air Pollution Unit, United Kingdom Department of Health*

 Cameographics, Union ME, Compositor. Printed by Recycled Paper Printing, Boston MA.
Library of Congress Catalog Number for the HEI Report Series: WA 754 R432.
Cover paper: made with 50% recycled content, of which 15% is post-consumer waste; free of acid and elemental chlorine.
Text paper: made from 100% post-consumer waste; acid free; no chlorine used in processing. The book is printed wtih soy-based inks and is of permanent archival quality.

PROJECT REPORT

Internet-Based Health and Air Pollution Surveillance System

Scott L Zeger, Aidan McDermott, Francesca Dominici, Roger Peng, and Jonathan Samet

INTRODUCTION

The Internet-Based Health and Air Pollution Surveillance System (iHAPSS*) provides researchers with publicly available data, software, and documents for conducting time-series studies on acute air pollution exposure and daily mortality in US urban communities. iHAPSS grew out of the HEI-funded National Morbidity, Mortality, and Air Pollution Study (NMMAPS) (Samet et al 2000a,b; Daniels et al 2004; Dominici et al 2005), a cooperative effort between the Bloomberg School of Hygiene and Public Health at the Johns Hopkins University and the Harvard School of Public Health. NMMAPS scientists studied how particulate and other air pollution might be associated with daily variations in hospitalizations and mortality; these analyses controlled for the confounding effects of season, weather, and other factors. The US Environmental Protection Agency relied on the NMMAPS results in its 2004–2005 review of National Ambient Air Quality standards for particulate and ozone pollution.

Time-series studies such as NMMAPS compare daily death rates with daily air pollution levels within the same population; by studying individual cities, they control for possible differences in unmeasured confounding variables among communities. These studies are critical for establishing whether air pollution at current levels in US cities causes premature death. Time-series studies are subject to possible confounding, however, by season (eg, influenza in winter) and weather; the collective influence of these two factors on mortality is an order of magnitude greater than the effect of particulate air pollution. Because of this, large multi-city data sets and sophisticated statistical methods and software are necessary for drawing valid inferences about the effects of air pollution.

Given the seriousness of the public health problem, the potential costs of new regulations, the need to rely upon observational rather than experimental human exposure studies, and the complexity of the data and analytic methods of inference, NMMAPS results have been challenged in public regulatory settings and in scientific publications (National Research Council [US] 2001; Health Effects Institute 2003). To address major concerns, NMMAPS investigators have conducted and published further analyses (Peng et al 2005) and have provided NMMAPS data and methods to others to do the same. However, NMMAPS funding was not intended to cover handling multiple requests for data and software; such requests would limit the time and resources for additional analyses.

We created the iHAPSS project in 2002 in order to make the data, statistical methods, and statistical software used in the NMMAPS mortality analyses available on the internet. In this way, others could check the data sets, reproduce NMMAPS results, conduct original analyses with our methods and software, or modify the methods to conduct novel analyses of these data. iHAPSS is an example of what geophysicists call reproducible research (Buckheit and Donoho 1995), an application of literate programming (Knuth 1992; Ramsay 1994).

The term *surveillance system* in the iHAPSS acronym is unconventional but chosen purposefully. Environmental health surveillance involves monitoring routinely collected exposure and health outcome data to detect public health risks. iHAPSS integrates billions of data bytes from four federal government agencies to help answer the question of whether air pollution exposures—at their current levels—cause premature disease and death.

This document includes information about:

- iHAPSS implementation, including the data, software, and analytic results it offers the public;
- iHAPSS users in a recent one-month period;
- suggestions for future improvements derived from a users' group meeting and our experience to date; and

* A list of abbreviations and other terms appears at the end of the Project Report.

This Project Report is one part of Health Effects Institute Communication 12, which also includes Comments on the project by members of the Health Research and Review Committees and other experts. Correspondence concerning the Project Report may be addressed to Dr Scott L Zeger, Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, 615 North Wolfe Street, Room E3132, Baltimore MD 21205-2179.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award R82811201 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

- iHAPSS as an example of reproducible epidemiologic research and the importance of expanding its use in environmental epidemiology.

To understand iHAPSS in its current form, its potential for the future, and the importance of reproducible epidemiologic research, please visit the website at *www.ihapss.jhsph.edu/* (Johns Hopkins Bloomberg School of Public Health 2005).

iHAPSS WEBSITE

The iHAPSS site was designed to maximize its utility to scientists who seek access to the data, methods, and software for conducting their own time-series studies of the association of daily pollution and mortality in US urban centers. Therefore, we made the maximum amount of data available, in a form that is easy to access and use, and provided the methods and software needed. Content is appropriate for users with substantial statistical expertise, whether statisticians or epidemiologists. The website was not specifically designed for the broader community of environmental scientists and regulators.

This section summarizes the major components available on the website and gives addresses of the appropriate pages for the reader who wants to review the site while reading this document. The initial iHAPSS page is pictured in Figure 1.

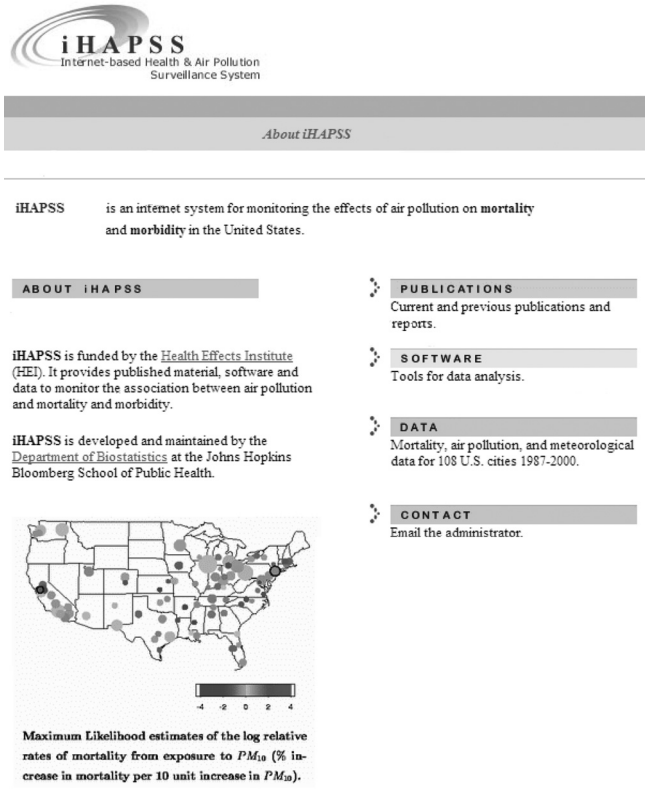

Figure 1. iHAPSS home page (*www.ihapss.jhsph.edu/*).

DATA INCLUDED IN iHAPSS

The iHAPSS website (*www.ihapss.jhsph.edu/data/*) provides daily time-series data from January 1, 1987 through December 31, 2000 (5114 days) on air pollution, weather, and mortality for the 108 US urban centers shown in Figure 2. The core data for each city comprise

- mortality counts by cause: total nonaccidental, cardiovascular, respiratory, chronic obstructive pulmonary disease, pneumonia, and accidental;
- pollutants: particulate matter (PM) less than 10 μm in aerodynamic diameter (PM_{10}), PM less that 2.5 μm ($PM_{2.5}$), SO_2, NO_2, O_3, and CO; and
- weather: temperature, dew point, and relative humidity.

We have compiled these data (hereafter referred to as iHAPSS data) from the National Center for Health Statistics, the US Environmental Protection Agency, the National Oceanic and Atmospheric Agency, and the US Census Bureau. We posted them into files organized by urban center (*www.ihapss.jhsph.edu/data/NMMAPS/descriptives/*). The pollution data for New York City are displayed in Figure 3 as an illustration. Substantial preprocessing was performed on the information accessed from federal databases in preparation for time-series studies. For example, for each urban center, we created a single PM_{10} time series from daily records available from all monitors in the counties comprising the urban center. We preprocessed that series to remove monitor-specific drift and to avoid outlying observations.

We have chosen R (*www.r-project.org*) for NMMAPS statistical analyses for several reasons. It is a free, open-source system that gives users complete access to all statistical routines. It is a language for data analysis that facilitates creation of specialized programs for novel analyses. Its facility for creating and disseminating new packages is unique among statistical software systems. It is highly graphical, which enables more effective displays of data and results. Finally, R has emerged as the standard software among research biostatisticians around the world.

Users of our statistical routines require the iHAPSS data in R format. Hence, the R package provides the core data for all 108 cities. It provides a number of utilities for abstracting and processing the data in preparation for statistical analysis. It also facilitates the kind of multi-city

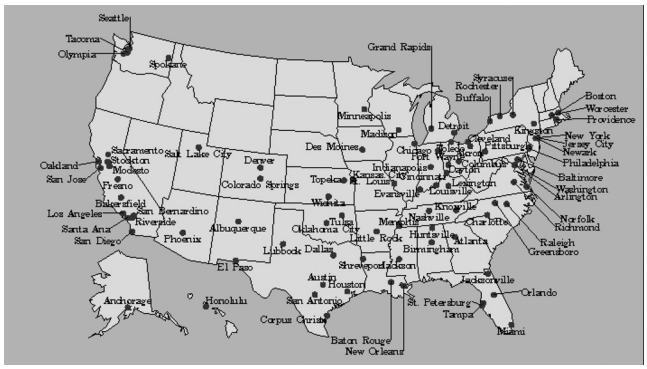

Figure 2. The 108 cities for which NMMAPS data are currently available from the iHAPSS website.

analyses conducted in the original NMMAPS project. Details about how to access and use the R data set are provided at *www.ihapss.jhsph.edu/data/NMMAPS/R/* and in Appendix B.

New York

Variable	N	Mean	Std Dev	Minimum	10th Pctl	Median	90th Pctl	Maximum
co	5111	1976.64	548.61	272.07	1335.68	1926.58	2664.10	5201.01
no2	5044	34.81	10.11	-0.76	23.59	33.12	48.33	86.06
so2	5111	12.77	8.22	0.22	5.04	10.60	23.43	76.48
o3	5113	19.59	11.06	-2.48	7.24	17.69	34.08	80.89
pm25	973	15.75	11.21	-16.43	5.01	13.00	30.37	73.89
pm10	864	27.44	12.71	1.58	15.53	24.33	42.82	127.55

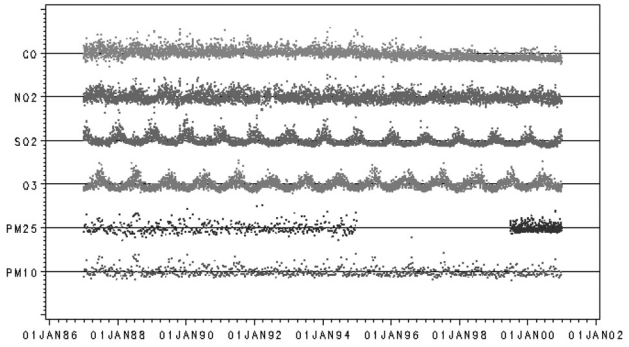

Figure 3. Pollution data available from the iHAPSS website for New York City (and each of the other 107 cities in the database).

STATISTICAL METHODS AND SOFTWARE

The statistical approach used in NMMAPS analyses has two stages (*www.ihapss.jhsph.edu/software/*). First, a log-linear regression model is applied to the time-series data for each city to estimate the city-specific relative risk of mortality per unit of change in air pollution while controlling for weather and time trends. Second, the city-specific relative risk estimates are pooled (smoothed) to obtain national and regional average relative risks and improved city-specific estimates. The improvement is achieved by borrowing strength across neighboring cities to overcome the substantial statistical noise in the values of the naïve estimates obtained from the first stage. (For examples of these methods see Samet et al 2000a; Bell et al 2004; Dominici et al 2004.)

The statistical methods applied in stages 1 and 2 are too complex to be fully documented in this paper or in any of the NMMAPS publications. Because the relative risk of mortality is small compared with the size of the possible confounding effects of season and weather, the results for any one city can be sensitive to modeling choices. Hence, it is essential that the exact procedures used by NMMAPS investigators be available to other scientists and policy analysts. This can only be done by providing the software used by the NMMAPS investigators.

We used the vignette system in R to disseminate and document the software. Included with the NMMAPSdata R Package (Appendix B) is an overview of the software package and the data. Examples of how to use the package are included, in which standard text is interspersed among the R code. Using the vignette system in R, the user can easily extract these code examples and run them separately. Another advantage of using R for distributing data is the strict quality control standard imposed by the R software for documentation and vignettes. Hence, the code included with any documentation is guaranteed to run.

In addition to providing the specific software used by NMMAPS investigators, we also provide a weblink to R for those who want to conduct analyses without establishing an R package on their own system. This allows web visitors to customize and run analyses with data existing on the website or with their own data. This had been envisioned as part of Phase II of the iHAPSS project and is introduced here to test its feasibility.

PUBLISHED PAPERS

In addition to providing the data and software, integrated via vignettes as described above, the website is currently used to disseminate papers produced by the NMMAPS study group *(www.ihapss.jhsph.edu/publications/)*.

WEBSITE AND DATABASE DESIGN

Equipment

The website is managed on two Dell Precision 340, 2.54 GHz computers with 1 GB of RAM each (Dell Inc, Round Rock TX). One computer has been configured to be the webserver and the other maintains the development environment. The webserver currently runs under Linux (7.3) (*www.linux.org*) to maximize resources. The development computer is configured for Microsoft Windows (Microsoft Corp, Redmond WA) as well as for Linux to exploit web-development software available for Microsoft Windows.

Webserver Software

Apache (1.3.23) (*www.apache.org*) is the webserver package, PostgreSQL (7.2.3) (*www.postgresql.org*) is used to manage database administration, and R (1.6) is our chief statistical package.

Data

We obtained and cleaned meteorologic, pollution, and mortality data from the US government sources named above. We downloaded and then preprocessed the data using SAS macros (SAS Institute, Cary NC) available in the software section on the iHAPSS website. Currently the website contains the following data:

- meteorologic (average daily temperature and dew point temperature through December 31, 2001);
- pollution (PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, CO through December 31, 2001); and
- mortality counts by three age groups (through December 31, 2000).

The National Center for Health Statistics recently denied our request to update the mortality statistics through December 31, 2001, citing new policies on data security. We continue to negotiate with them to address their security concerns and obtain the data.

iHAPSS USERS

OVERVIEW

The iHAPSS website is open to the public without any registration requirement. This is both a strength and a limitation. Although it may encourage visits by those who wish to remain anonymous, it limits our ability to identify the user population. Such identification would permit site developers to conduct periodic surveys to find out which components of the current site are most useful to visitors and to identify the need for new components. An important question for the future is whether registration should be required before one can download data, software, or papers.

We collected data on the visitors from their URLs to get an overview of the number and types of visits from July 12 through August 8, 2004. The iHAPSS server received 30,000 requests with 783 unique IP (internet protocol) addresses; 600 of those were traceable. We determined that 346 addresses were not spiders (computer programs that search the internet to update search engines) or other automatic queries. Thus, the site received about 11 unique visitors per day during that time. The visitors were from more than 10 countries including the United States, Canada, United Kingdom, France, South Africa, Netherlands, Australia, New Zealand, Germany, Japan, Italy, Spain, Brazil, and Denmark. They represented several US federal and state agencies including the US National Institutes of Health, US Centers for Disease Control and Prevention, and state health departments from New York, New Jersey, and Colorado.

USERS' GROUP MEETING

We convened a group of scientists and environmental policy experts (Appendix A) during a one-day meeting to review the current iHAPSS and to make recommendations about its future development. The discussion was organized around four questions; the participants' comments are summarized here.

Who Is the Potential Audience or Constituency for iHAPSS?

The users' group thought that the future audience for iHAPSS, like its current one, will likely be environmental scientists and policy analysts with at least a Masters-level expertise in statistics. iHAPSS enables such persons to engage in statistical modeling of air pollution and mortality data without expending the substantial start-up costs necessary to produce the time-series data sets and statistical programs. It was not designed for people with limited statistical expertise.

During the original design of the iHAPSS project, we envisioned a second stage in which the site might be redesigned so that less technically-oriented persons could conduct rudimentary time-series analyses. The users' group thought this strategy unlikely to be successful and agreed that it should not be pursued.

How Can iHAPSS Better Meet Scientists' Needs to Access Air Pollution and Health Data?

The major issue was how to optimally compile and construct the iHAPSS database. Publicly available data are abstracted from several governmental databases and substantially preprocessed to meet the specific needs of the NMMAPS investigators. The users' group thought that future users would want to compile and construct the database in different ways. For example, the currently posted daily iHAPPS ozone data are daily means. Some users might prefer the 3- or 8-hour maximum. One strategy would be to post the hourly data from which the daily values were aggregated. Currently, the software provided is only designed to recreate NMMAPS data sets from public sources. The group wanted others to be able to redo the analyses with either the original or the recreated data set.

The group thought that the iHAPSS utility would be diminished if it were simply a second copy of the many databases from which iHAPSS abstracts data. The topic remains open for further consideration.

A second important issue was whether to expand iHAPSS beyond the NMMAPS database. For example, Johns Hopkins investigators are using daily time series and annual total mortality counts derived from the Medicare Cohort Study to simultaneously estimate acute and chronic exposure mortality relative risks. The number of persons exposed and the number of deaths in the 300 largest counties where air pollution data are consistently available are listed below. Permission from the Center for Medicare and Medicaid Services is required before these data can be made public.

- 2000; 1,022,000 deaths; 19,680,000 at risk
- 2001; 1,027,000 deaths; 19,767,000 at risk
- 2002; 1,033,000 deaths; 19,838,000 at risk

How Might iHAPSS Be Developed as a Public Health Tool for Surveillance of the Health Effects of Air Pollution?

Whether and how air pollution causes morbidity and mortality are questions that will be with us for years. Roger Peng presented recent findings about seasonal and geographic variations in the relative risk of mortality from pollution. These findings raised numerous questions about the underlying mechanisms involved in these variations (Peng et al 2005; *www.bepress.com/jhubiostat/paper41/*). Regular (eg, yearly) iHAPSS updates are needed to allow researchers to monitor whether government regulations that change PM characteristics (size, composition) and toxicity change the PM health effects.

The users' group also discussed how a national analysis like NMMAPS could be important to local regulatory and public health decisions. Historically, local or state health departments use only their own data to monitor exposures and health effects in their populations. With the availability of iHAPSS data and methods, it is now possible to substantially improve the estimates for a particular region by borrowing strength across neighboring and other similar regions. The technology for doing so is not well understood nor is it used by environmental regulators. iHAPSS represents an opportunity to begin the education process that could improve local decision making.

How Might iHAPSS Be Developed and Promoted as a Model for Replication of Findings in Epidemiology?

Analyses like NMMAPS are sufficiently complex that they cannot be fully documented and described in conventional publications. If easy access to data and methods is provided, independent investigators can confirm major findings of the original analyses or offer alternate explanations if they arrive at different results.

Reproducing study results is not the same as replicating a study, however. To reproduce study results, investigators use the same data and methods from one study to document that they can produce the same results as the original investigators: iHAPSS allows the original NMMAPS results to be reproduced. Study replication requires an independent data set from a different and distinct population. For example, the HEI-funded APHENA project being conducted by Samet and colleagues as discussed in the HEI 2005–2010 Strategic Plan (2005) was designed to replicate the NMMAPS study (Samet et al 2000a,b) in populations across the US, Canada, and Europe.

The group suggested that iHAPSS could usefully be promoted as an example of the tools available to allow reproduction and validation of results in epidemiologic research. We close with a brief discussion of the concept of reproducible research.

GENERAL DISCUSSION OF REPRODUCIBLE EPIDEMIOLOGIC RESEARCH

A decade ago, Taubes (1995) questioned the reliability of observational epidemiologic studies for quantifying health effects of risk factors such as second-hand smoke, air pollution, and diet. A number of trends contribute to discounting current epidemiologic findings. First, the signal-to-noise ratio in more recent studies tends to be smaller than it was in decades past. Major diseases have well established large relative risk factors, for example smoking, socioeconomic status, family history, and obesity. More recent investigations tend to target factors with smaller relative risks that are more easily confounded. For example, the NMMAPS team estimated the relative risk of increased mortality in the United States to be 1.005 per 10 ppb of 24 hour ozone (Bell et al 2004). While the relative risk is small, it translates into thousands of excess deaths per year given the universality of ozone exposure. Nevertheless, the potential for unexplained confounding is ever-present for such a small risk ratio.

Another factor is the explosion of new biologic measurements, products of the twin information and biotechnology revolutions. We can now quantify DNA sequences, single nucleotide polymorphisms, and gene and protein expression. We can image the structure and function of the brain and other organs. We quantify diet with lengthy dietary-recall questionnaires. We quantify disease symptoms and health conditions using multi-item instruments.

These modern measurements are used both as outcomes and risk factors in epidemiologic studies. They are inherently high-dimensional and subject to considerable natural variability and measurement error.

The new measurement technologies have implications for epidemiologic studies; they obviously open exciting new opportunities. However, because they are high-dimensional, the potential for identifying spurious associations between a selected subset of risk factors or health outcomes is increased compared with analyses that use a smaller number of variables. For example, if we search for genes that interact with diet to cause disease using gene expression arrays, there are tens of thousands of potential risk factors to consider.

A related trend is the focus of epidemiologic studies on interactions among multiple risk factors, for example gene–environment interactions. The study of interactions requires study populations to be partitioned into much smaller jointly-exposed groups or into subsets with predispositions to increased risk, thereby increasing the statistical noise—even in the presence of larger signals for the subgroup.

Another factor contributing to the potential for false positive epidemiologic findings is the widespread availability of statistical and computing technology. It is routine to engage in sophisticated searches across a large number of variables for associations of potential scientific interest. As the number of covariables measured increases, so do the degrees of freedom for influencing the association between a particular risk factor and outcome and for identifying subgroups in which the association is large.

Finally, smaller relative risks and novel measurement technologies are by necessity leading to larger, longer, and more expensive studies. Hence, the pressure to produce and publish results is magnified. This too contributes to the potential for false positive findings.

Of course, the very trends identified above as contributing to the potential for an increase in the occurrence of spurious findings can, and in many cases have, dramatically increased the power and precision of epidemiologic research. The information and biotechnology revolutions have advanced our understanding of disease mechanisms. Because prior biologic knowledge is improved, epidemiologic studies can test more targeted, mechanism-driven hypotheses. Because there are more direct biologic measurements, these hypotheses can be addressed with greater precision. Because modern computing makes the organization, management, and analysis of large databases possible, we can look further and wider for systematic patterns indicative of the health effects of risk factors.

The net effect is that the validity of epidemiologic studies is increasingly dependent upon more sophisticated and complex measurement technologies, databases, and statistical analyses. Conducting epidemiologic research requires increased biologic understanding and statistical rigor to achieve the potential increases in precision and to avoid the pitfalls associated with smaller targets, higher-dimensional measurements, and misapplied statistical and computing power.

Making data, methods and software available to scientific colleagues and critics is an essential first step toward achieving statistical rigor. Data sets and analyses for projects like NMAPPS are sufficiently complex that they can not be reproduced or easily extended without the complete sharing of these components. It is simply not possible to sufficiently explain the details required to assure reproducibility. iHAPSS is one example of a research area with important policy implications that can provide that first step toward reproducible epidemiologic research.

ACKNOWLEDGMENTS

The authors acknowledge partial support for their work provided by the National Institute of Environmental Health Sciences (grants R01ES012054 and P30 ES 03819). They thank Ms Debra Moffitt for her administrative support and the members of the Johns Hopkins Environmental Biostatistics and Epidemiology Working Group for discussions of this project.

REFERENCES

Bell M, Samet JM, McDermott A, Zeger SL, Dominici F. 2004. Ozone and mortality in 95 US urban communities from 1987 to 2000. JAMA 292:2372–2378.

Buckheit JB, Donoho DL. 1995. Wavelab and Reproducible Research. Stanford University, Stanford CA. Available from *www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf* Accessed May 9, 2006.

Daniels MJ, Dominici F, Zeger SL, Samet JM. 2004. The National Morbidity, Mortality, and Air Pollution Study: Part III: PM_{10} Concentration-Response Curves and Threshold for the 20 Largest US Cities. Research Report 94. Health Effects Institute, Boston MA.

Dominici F, McDermott A, Hastie T. 2004. Improved semiparametric time series models of air pollution and mortality. J Am Stat Assoc 468:938–948.

Dominici F, Zanobetti A, Zeger SL, Schwartz J, Samet JM. 2005. The National Morbidity, Mortality, and Air Pollution Study: Part IV. Hierarchical Bivariate Time Series Models:

A Combined Analysis of PM_{10} Effects on Hospitalization and Mortality. Research Report 94. Health Effects Institute, Boston MA.

Health Effects Institute. 2003. Commentary on revised analyses of selected studies. A special panel of the Health Review Committee. In: Revised Analyses of Time-Series Studies of Air Pollution and Health. Special Report. Health Effects Institute, Boston MA.

Health Effects Institute. 2005. HEI Strategic Plan for Understanding Health Effects of Air Pollution 2005–2010. Health Effects Institute, Boston MA.

Johns Hopkins Bloomberg School of Public Health. 2005. Internet-based Health and Air Pollution Surveillance System. *www.ihapss.jhsph.edu*. Last updated March 19, 2005. Accessed May 9, 2006.

Knuth D. 1992. Literate Programming. Center for the Study of Language and Information Lecture Notes, Vol 27. Stanford University, Stanford CA.

National Research Council (US). 2001. Research Priorities for Airborne Particulate Matter: Part III. Early Research Progress. National Academy Press, Washington DC.

Peng R, Dominici F, Pastor-Barriuso R, Zeger SL, Samet JM. 2005. Seasonal analyses of air pollution and mortality in 100 US cities. Am J Epidemiol 161(6):585–594.

Ramsey D. 1994. Literate programming simplified. IEEE Software 11(5):97–105.

Samet JM, Dominic F, Zeger SL, Schwartz J, Dockery DW. 2000a. The National Morbidity, Mortality, and Air Pollution Study: Part I. Methods and Methodologic Issues. Research Report 94. Health Effects Institute, Cambridge MA.

Samet JM, Zeger SL, Dominici F, Curriero F, Coursac I, Dockery DW, Schwartz J, Zanobetti A. 2000b. The National Morbidity, Mortality, and Air Pollution Study: Part II. Morbidity and Mortality from Air Pollution in the United States. Research Report 94. Health Effects Institute, Cambridge MA.

Taubes G. 1995. Epidemiology faces its limits. Science 269(5221):164–169.

APPENDIX A: Users' Group Participants

Investigators

Scott Zeger, Principal Investigator, Professor and Chair, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University (BSPH)

Francesca Dominici, Coinvestigator, Associate Professor, Department of Biostatistics, BSPH

Aidan McDermott, Coinvestigator, Assistant Scientist, Department of Biostatistics, BSPH

Other Participants

Tim Buckley, Associate Professor, Department of Environmental Health Sciences, BSPH

Alison Geyh, Assistant Professor, Department of Environmental Health Sciences, BSPH

Barbara Glenn, Scientist, National Center for Environmental Research, EPA

Dennis Kahlbaum, Consulting Meteorologist and Senior Analyst, Air Improvement Resource, Inc

Thomas Lumley, Associate Professor, Department of Biostatistics, University of Washington, Seattle

Sumi Mehta, Staff Scientist, Health Effects Institute

Robert O'Keefe, Vice President, Health Effects Institute

Roger Peng, Assistant Professor, Department of Biostatistics, BSPH

Paul Rathouz, Assistant Professor, Department of Health Studies, University of Chicago

Howard Rockette, Professor and Chair, Department of Biostatistics, University of Pittsburgh; HEI Research Committee

Margaret Round, Senior Air Toxics Program Analyst, Northeast States for Coordinated Air Use Management (NESCAUM); now at Massachusetts Department of Public Health

Michael Stein, Professor, Department of Statistics, University of Chicago

Leah Welty, Assistant Professor, Department of Biostatistics, Northwestern University Medical School

Ronald White, Associate Scientist, Department of Epidemiology, BSPH

APPENDICES AVAILABLE ON REQUEST

APPENDIX B. NMMAPSdata R Package

This package is available at *www.ihapss.jhsph.edu/data/NMMAPS/R/* or may be downloaded from the HEI website, *www.healtheffects.org*. If you need assistance obtaining this information, contact HEI at Health Effects Institute, Charlestown Navy Yard, 120 Second Avenue, Boston MA 02129-4533, +1-617-886-9330, fax +1-617-886-9335, or email (*pubs@healtheffects.org*). Please give (1) the first author, full title, and number of the Communication and (2) title of appendix requested.

ABOUT THE AUTHORS

Scott L Zeger is the Hurley-Dorrier Professor and Chair of the Department of Biostatistics at the Johns Hopkins University Bloomberg School of Public Health. He earned his PhD in statistics from Princeton University and conducts research on statistical methods for longitudinal and time-series data with applications to environmental and biomedical data.

Aidan McDermott is Associate Scientist in the Department of Biostatistics at the Johns Hopkins University Bloomberg School of Public Health. He has a PhD in mathematics from the National University of Ireland, Galway. He is a specialist in the design, management, and analysis of complex databases. Dr McDermott's research is in the environmental and biomedical sciences.

Francesca Dominici is Associate Professor in the Department of Biostatistics at the Johns Hopkins University Bloomberg School of Public Health. She earned a PhD from University of Padua, Italy, in statistics. Her research is on the health effects of air pollution and more generally on statistical methods for biomedical and epidemiologic research.

Roger Peng is Assistant Professor in the Department of Biostatistics at the Johns Hopkins University Bloomberg School of Public Health. Dr Peng has a PhD in statistics from the University of California at Los Angeles. His research is on statistical methods for spatial time-series data with application to environmental and epidemiologic data.

Jonathan Samet is Professor and Chair of the Department of Epidemiology at the Johns Hopkins University Bloomberg School of Public Health. He earned his MD at the University of Rochester and his Masters in Epidemiology from Harvard University. Dr Samet conducts epidemiologic studies of the health effects of environmental exposures including air pollution and smoking.

OTHER PUBLICATIONS RESULTING FROM THIS PROJECT

Peng RD, Dominici F, Zeger SL. 2006. Reproducible epidemiologic research. Am J Epidemiol 163(9):783–789.

ABBREVIATIONS AND OTHER TERMS

EPA	US Environmental Protection Agency
GAM	generalized additive model
iHAPSS	Internet-Based Health and Air Pollution Surveillance System
NMMAPS	National Morbidity, Mortality, and Air Pollution Study
PM_{10}	particulate matter less than 10 μm in aerodynamic diameter
$PM_{2.5}$	particulate matter less than 2.5 μm in aerodynamic diameter
OMB	US Office of Management and Budget
R	statistical software package chosen for NMMAPS analyses
FOIA	Freedom of Information Act

HEI Communication 12, *Internet-Based Health and Air Pollution Surveillance System*, SL Zeger et al

INTRODUCTION

This document describes a project by Dr Scott Zeger and colleagues of the Johns Hopkins Bloomberg School of Public Health that was funded by HEI with the purpose of making data and software from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS*; see sidebar next page) available to a wide range of interested parties. It provides a brief overview of the project and comments from some members of the HEI Health Research and Review Committees and other experts.†

HEI has always recognized that to be credible, the science it funds and the data underlying the science must be transparent to other interested parties. In response to increased interest in its science, in particular the Reanalysis of the Harvard Six Cities and American Cancer Society Studies (Krewski et al 2000), in the late 1990s HEI adopted its first policy to ensure public access to data from studies it funds. Around that time, public access became a higher priority for the broader scientific community because of (1) the National Academy of Sciences recommendations on data issues in the natural sciences (National Research Council 1997) and (2) the Shelby Amendment (a provision inserted in the fiscal year 1999 Omnibus Appropriations Bill to allow broader access to federally funded research data). After the federal regulations for implementing the Shelby Amendment were issued, HEI formalized its Data Access Policy (Appendix A) to ensure consistency with provisions in the Shelby Amendment. Beyond that, and in the spirit of transparency and fostering data access, HEI sought to identify more proactive ways to make data available to the public, especially for high-impact projects like NMMAPS.

Increasing access to data enhances credibility by providing transparency about the research process and data interpretation. It also provides an opportunity for other researchers to validate and reanalyze the results and to perform additional analyses. HEI has extensive experience conducting reanalyses of studies that have regulatory importance: for example, the Particle Epidemiology Evaluation Project (Samet et al 1995, 1997), the studies of railroad workers that evaluated their exposure to diesel emissions and incidence of lung cancer (Health Effects Institute 1999), the Reanalysis of the Harvard Six Cities and American Cancer Society Studies (Krewski et al 2000), and most recently the Revised Analyses of Time-Series Studies of Air Pollution and Health (Health Effects Institute 2003).

Reproducibility of results is an integral and important part of the scientific process. Both reproducibility (using the same data set) and replication (using data from a different study population) of epidemiologic studies have received renewed attention (Neutra et al 2006; Peng et al 2006). They reemphasize the importance of conducting reanalyses and making data publicly available.

Because of the importance of NMMAPS in science and regulatory decisions, research groups interested in conducting additional data analyses have requested access to NMMAPS data. In response, Dr Zeger and colleagues submitted a proposal to create a website that would provide electronic access to NMMAPS data and to the software used for the data analyses. This approach was intended to facilitate study replication and new analyses. It would also reduce the burden on the NMMAPS investigators to respond to numerous individual requests for information. HEI was interested in funding the project because it addressed two important science and policy issues: (1) facilitating public access to data from scientific studies that figure prominently in regulatory decisions; and (2) providing access to regularly updated databases that could allow ongoing surveillance of the health effects of air pollution as ambient concentrations change over time in response to regulatory activity or other causes.

* A list of abbreviations and other terms appears at the end of the Project Report.

†Dr Zeger's 2-year project, "Internet-Based Health and Air Pollution Surveillance System (iHAPSS)," began in April 2002. Total expenditures were $488,330. A draft Project Report from Zeger and colleagues was received in January 2005 and accepted for publication in June 2005. During the evaluation process, comments were provided by some members of the HEI Health Research and Review Committees and by other experts. The investigators had the opportunity to exchange comments and to clarify issues in the Project Report and these Comments.

‡The Shelby Amendment requires that federally supported research be available under the Freedom of Information Act (FOIA) if the research is used to develop a federal agency action that has the force and effect of law. A brief history of the Shelby Amendment is provided on the AAAS website (American Association for the Advancement of Science 2005).

This document has not been reviewed by public or private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views of these parties, and no endorsements by them should be inferred.

The National Morbidity, Mortality, and Air Pollution Study

Over the past decade, many researchers have used time-series studies to evaluate the association between daily changes in the particulate matter (PM) concentrations in ambient air and the daily morbidity and mortality for individual cities. HEI funded NMMAPS to address concerns about bias in the selection of cities included in time-series analyses by including data from the 90 largest US cities that had PM data. NMMAPS was conducted by Dr Jonathan Samet and colleagues at Johns Hopkins University in collaboration with investigators at Harvard University and was published as HEI Research Report 94 (Samet et al 2000a,b; Daniels et al 2004; Dominici et al 2005); additional development of methods and statistical models is described in HEI Research Report 123 (Dominici 2004).

NMMAPS employed national databases on air pollution and health outcomes to evaluate the acute effects of air pollution and applied sophisticated statistical approaches, some of which were developed specifically for the study. The study received national and international attention because it addressed some uncertainties regarding the association between PM and daily mortality and determined the effects of other pollutants on this association. Its results have been used by the US Environmental Protection Agency (EPA) in the setting of PM standards.

NMMAPS provides estimates of the effects of PM_{10} (particulate matter less than 10 μm in aerodynamic diameter) and gaseous air pollutants on daily mortality in the 90 largest US urban centers with a combined population of 94 million people. In a parallel analysis, investigators at the Harvard School of Public Health investigated the effect of air pollution on daily hospital admissions of elderly individuals for cardiovascular disease, chronic obstructive lung disease, and pneumonia in 14 cities. This comprehensive multisite design addressed many of the limitations of earlier single-city analyses by using a unified analytic approach to examine the effects of PM_{10} and other pollutants in a large number of cities that span the continental United States and vary widely in their average levels of criteria air pollutants.

After completion of the project, but during continuing analyses, the NMMAPS investigators identified statistical issues that indicated a problem with using generalized additive models (GAMs) to account for time-varying factors, such as temperature and humidity. Although many methods can be used to adjust for time-varying factors, GAMs have become the favored method in recent years. The NMMAPS investigators discovered that part of the programming in the S-Plus statistical software was inappropriate to analyze such data, because under some conditions the iterative process to obtain effects estimates does not converge to the true estimate of the regression coefficients. In addition, investigators at Health Canada discovered that under certain conditions the GAM software resulted in underestimates of the standard errors. A large effort was undertaken by the NMMAPS and other investigators, the EPA, and HEI to address the statistical issues. Revised and new analyses of NMMAPS and 21 reports on results from other time-series studies were published in HEI's Special Report, Revised Analyses of Time-Series Studies of Air Pollution and Health (Health Effects Institute 2003).

The HEI Special Panel that reviewed the revised analyses noted that neither the appropriate degree of control for time nor appropriate specification of the effects of meteorologic factors had been determined. The Panel recommended that future efforts should explore how much to control for time in time-series analyses. The Panel commented that the survival of residual time effects in these studies indicates a need to measure other potential risk factors, such as those related to weather, and take them into account in the analytic models. The Panel also recommended exploring the sensitivity of these studies to a wide range of alternative degrees of smoothing and to alternative specifications of meteorologic factors. In response to these recommendations, HEI initiated new research on comparative and alternative methodological approaches to time-series analysis. It is envisioned that the iHAPSS website will be a useful tool to extend the ability of qualified researchers to conduct additional analyses and explore different methods using the NMMAPS data.

The project had the following objectives:

- provide an opportunity for other researchers to reproduce the original analyses by including the data analysis tools with the data; and
- make it possible for other researchers to conduct additional analyses of the NMMAPS data that are tailored to their individual interests by providing flexibility in data format, data selection, and software.

The project was also anticipated to reduce demands on the NMMAPS investigators to respond to many requests for data and increase overall study transparency. This model could be applied to other research expected to yeild results of high interest for regulatory and public health purposes.

PROJECT DESCRIPTION

Dr Zeger and colleagues submitted a proposal with the objective to "develop an internet-based statistical system for assessing the effects of air pollution on daily mortality and morbidity in United States cities". The proposed project contained two phases:

1. Create an internet site to disseminate data, statistical software, and regularly updated results from NMMAPS.
2. Design and implement a web-based interactive system to enable users to conduct their own analyses using NMMAPS methods.

As a result of work performed in Phase 1, users would be able to:

- obtain time-series data for the cities and time periods of interest;
- obtain time-series data on mortality, air pollution, and confounding variables;
- conduct statistical analyses of these data to estimate relative risks;
- optimally pool results across cities; and
- display geographic patterns in relative risks for regions of interest.

Phase 2 would add access to additional monitoring data from the US Environmental Protection Agency (EPA) and other sources; the Internet-Based Health and Air Pollution Surveillance System (iHAPSS) website would not be a repository for these data but would facilitate the user's ability to obtain the requisite data from the publicly available data systems. HEI funded Phase 1 as a 2-year project. The project started in April 2002.

The resulting website (*www.ihapss.jhsph.edu*) provides daily time-series data on air pollution, weather, and mortality for 108 US urban centers for years 1987–2000. Data were compiled from the National Center for Health Statistics, the EPA, the National Oceanic and Atmospheric Agency, and the US Census Bureau. They are made available as text files and as a data package for use with the statistical software program R. The website also contains sections with software for data manipulation and analysis, an interface to R, and publications resulting from NMMAPS. Anyone can enter the website, although it was designed primarily for users with a background in statistics and epidemiology.

PROJECT EVALUATION

In the spring of 2004, as the project neared completion, the investigators and HEI assembled a group of sponsors and others from the community of expected users to assess the practical operation and ultimate utility of the website. A list of members and a description of the discussion and evaluation is included in the Project Report. After the project was completed, HEI approached this group and others in government, industry, and academia to provide written comments evaluating the website. Some members of both the HEI Health Research and Review Committees also provided written comments about the project. This document provides a compilation of the comments that were received; these are organized by a set of questions that were presented to the evaluators.[§]

COMMENTS ON THE PROJECT

1. *Has the iHAPPS website accomplished its goals?*

Because of the regulatory implications of the NMMAPS study, evaluators agreed this was an important project, especially as a prototype for the appropriate presentation of statistical analyses that play a role in major public policy decisions. Evaluators thought that the iHAPSS website has been a successful way to provide public access to detailed study information, including the actual NMMAPS data and software used to generate the results. In addition to summaries of results provided in several downloadable formats, the original articles and reports are also included, benefiting a broad community of researchers, policymakers, and other interested parties. The website also provides links to sources of air pollution and mortality data. Because of the structural organization of the internet, the website allows flexibility for a user to go as deeply into the level of detail and complexity of data and analyses as needed.

Most evaluators considered it important that researchers can access the tools used for data analysis. A website such as iHAPSS can provide this function, which is not possible in peer-reviewed journals with limited space for details about methods and software. Some evaluators reported that examining the software code provided a good way to understand the specific models and variables used, which allows users to carefully investigate the methods. In risk assessments for standard setting and other purposes, an increasingly important issue has been how much of an effect air pollution has on health; thus, questions have arisen about quantifying the health effects of air pollution and the nature of the exposure-response functions. A website can provide more detailed information than a journal article, such as how pollutant concentrations were averaged, the range of concentrations, the exact descriptions of the health outcomes and populations studied, and the

§ We use the term *evaluator* to indicate all individuals (those who participated in the original users' group, HEI Committee members, and others) who were asked to comment on the iHAPSS project.

types of statistical analyses. These and other details would be needed to perform further analyses or meta-analyses.

2. *What suggestions would you make to improve the approach (of making data and software available on a website)?*

Evaluators thought the iHAPSS website could potentially also provide an opportunity for ongoing surveillance of the health effects of air pollution as pollutant concentrations change over time due to air pollution regulations and other causes. However, this would require that data be updated regularly, which would need additional funding for the longer term.

In light of continued interest in NMMAPS data, especially in evaluating results using different methods, evaluators generally agreed it would be useful to maintain the current website for at least the short term.

One evaluator suggested that making the results more accessible to public health scientists at state agencies or in specific cities may be helpful because implementing air pollution standards occurs at the state level. In several states, environmental public health tracking programs (organized by the Centers for Disease Control and Prevention) are establishing surveillance programs that attempt to link environmental exposures to chronic diseases. NMMAPS could provide relevant time-series data for those states.

In addition, evaluators suggested ways to improve the website's visibility, one being to provide a link from the HEI website; users may not easily remember the acronym or exactly how they arrived at the website from other internet locations. Some evaluators mentioned that they had to repeatedly use an internet search engine to find the iHAPSS website. If similar projects are funded by HEI in the future, the HEI website could provide access to all such projects and to a single repository of papers describing research methods.

3. *Should HEI support similar efforts for studies of high regulatory importance and broad interest? If so, what could be some of the boundaries or restrictions on the kinds of projects?*

Evaluators were generally supportive of future projects, especially for multicity studies for which interest could be expected from diverse entities. Prerequisites for a project would be high regulatory importance, high data quality and experienced investigators who would maintain and update the website. The evaluators thought the iHAPSS project fit well with HEI's goal of communicating to the public and to scientific and regulatory communities its activities and research, especially those that are relevant to policy decisions.

Because of the costs incurred for such projects, some evaluators preferred funding additional projects rather than improving or expanding the iHAPSS website. One evaluator said it will be interesting to follow the emergence and development of similar websites, because increasingly higher demand for detailed information and increased levels of sophistication may result. Another suggested that such websites should indicate whether additional results or analyses may become available at a later time due to the ongoing nature of such projects.

On the other hand, one evaluator doubted that the iHAPSS project had added to the scientific knowledge and thought that reproducing analyses was not useful. This evaluator thought that, although it would be good to make the data assembly process transparent, it would be more informative if researchers tried to assemble the data sets from scratch rather than reanalyzing an existing data set. Making the data available without the software would have been a much cheaper option; the sophisticated data analysis capabilities were apparently underutilized (see below). This should be considered when designing future projects.

Some evaluators recommended expanding the website but most thought the needs involved in maintenance or expansion would be too costly. In addition, confidentiality issues about adding human subject data must be assessed carefully before deciding how to continue.

COMMENTS ON USING THE WEBSITE

1. *What were your goals as a user? Is the website user-friendly in general?*

Evaluators came from a variety of backgrounds, with some technical expertise, although not necessarily in epidemiology or statistics. Most evaluators had some understanding of the NMMAPS project and were interested in accessing the data, results, and summaries for specific pollutants or for specific cities; a minority were interested in performing new data analyses with the software. The website was considered user-friendly because they could easily find the information they wanted, such as city-specific estimates of relative risk. They also liked the maps and summary statistics, the collection of NMMAPS papers and reports, and information about problems with fitting generalized additive models (GAM) in the S-Plus statistical package that were first identified by the NMMAPS investigators and others (Health Effects Institute 2003; see the sidebar).

Overall, usage of the website has been modest, which is to be expected for a website that is so specialized. Zeger reported that the website received more extensive traffic after the GAM problems were identified because it provided software to calculate valid standard error estimates

in GAM (he reported an average of 11 visitors per day for August 2004).

2. *Could you access and manipulate the data and the software effectively? Was the information available in the format you desired and at the right level of detail? Are the on-line documentation and help functions adequate?*

Evaluators reported that they encountered no difficulties in downloading or manipulating the data, executing the examples provided, or exporting data into other software programs. They found the program to be flexible; a user is free to download data from any single city or set of cities. They appreciated the statistical NMMAPSdata R Package included with the software. The advantages they noted are (1) the ability to function on all major operating systems; and (2) that it contains both the data and the functions for manipulating the data to construct data sets for specific statistical analyses. Evaluators thought the documentation for these functions was adequate, provided the researcher had the expertise required to use the software. Some information, however, may be useful for persons with less technical expertise; for example, congressional science staff or city council assistants may benefit from the frequently asked questions and city-specific summaries even though they may not fully understand NMMAPS and its results.

The section containing the NMMAPSdata R Package is a major feature of the iHAPSS website. Evaluators described it as an impressive accomplishment and thought it could set a standard for making complex databases available to other researchers. The fact that it is an integrated data system that incorporates both data and modules to organize the data makes it powerful and saves time and effort for other researchers between downloading initial data and beginning statistical analyses. They considered the quality control of this section to be very good. The major drawback is that users need to be familiar with the R package to use its statistical functions, and experience with statistical software packages in general is required. However, users who are not familiar with the R package—and may experience difficulty using it—have the option to access the raw data instead, and apply their own preferred statistical package.

3. *Do you have suggestions for improving the current website?*

Evaluators thought that it may be useful to the general audience to provide more information and context about NMMAPS on the opening page, because many people in the public health field do not necessarily know the details about the study. Information could include summary conclusions with tables and maps, a brief description of the approach and methods, and caveats about the data and analyses.

More information could be added about design decisions the NMMAPS team made while constructing the database. For example, it would be helpful to read about how missing values were handled, how data from multiple monitors in a city were combined, how "cities" were defined (using political boundaries or a broader area), and how problems interpreting causes of death were handled.

The two sections entitled "How the mortality data were put together" and "How the pollution data were put together" provided nice graphic flowcharts for how raw data obtained from various national sources were transformed into the daily data sets used in the NMMAPS projects. These flowcharts could be useful to other researchers who may need to construct data sets, particularly if they want to construct variables in exactly the same way. For example, there are many ways to aggregate hourly ozone data from multiple sites at the daily and city level. A researcher might be interested in using same algorithm as was used in NMMAPS. The value of these two sections would be increased by interactively linking the SAS programs and the raw data sets to the graphics, so that a user could click on the graphic to view the appropriate data or program object.

In addition, it may be helpful to add more documentation on the overall structure of the R package, some basic terminology, and more detailed descriptions of the variables. For example: How were the "adjusted 3-day lag temperature" and "trimmed mean NO_2" computed? What is the difference between a database and a dataframe? What are the functions of loadCity, readCity, and attachCity? Do the coefficients in the Excel files describe results for "all ages" and "all years", or certain age classes and years?

More data export utilities could be added for those who would like to process the linked data sets in a different programming environment than the R package (such as Stata or SAS). Other evaluators suggested accessibility for other operating platforms or different internet browsers, specifically Macintosh operating systems and nonMicrosoft internet browsers.

If the database is regularly updated, it could be useful to include a suite of R functions that are "web-aware": that is, instead of downloading the entire NMMAPS database each time and accessing it locally, users' commands could be sent to the NMMAPS server and executed on the server; relevant results would be returned via the internet. In this way, the users would avoid downloading large data sets and be encouraged to make more frequent updates to their analyses, resulting in more accurate scientific results. In addition, it would be more user-friendly because data

download and database construction steps would be merged; the user would download only what was needed. The software could also contain a "check for updates" feature, so that analyses would not run unless the updates were included in the user's copy of the database.

It may be useful to develop an interface between the website and data output that ultimately links exposure to pollutants with type of disease, similar to the approach that has been used for the National-Scale Air Toxics Assessment (see *www.epa.gov/ttn/atw/nata/*).

SUMMARY OF EVALUATORS' COMMENTS

Most evaluators were satisfied that the website was user-friendly and provided easy access to the NMMAPS data. Some recommended adding more background information about NMMAPS for visitors who are not familiar with the study. Most appreciated the documentation and availability of the data, as well as links to related websites and the publications that resulted from NMMAPS.

In general, evaluators thought the goals of the iHAPSS website were appropriate and that the project had accomplished what it set out to do: make the data and software of NMMAPS available to a wider audience. Several evaluators were also supportive of similar future projects, provided that the data have the regulatory significance to warrant the effort and expense.

CONCLUSIONS

HEI supported the development of the iHAPSS website as part of its ongoing commitment to provide open public access to data from studies funded by the Institute, particularly studies of significant regulatory, scientific, and public health impact such as NMMAPS. Beyond providing open and transparent access to facilitate replication and validation efforts, HEI saw iHAPSS as an opportunity to facilitate new analyses of the NMMAPS data set, possibly adding to the scientific literature, while reducing the burden on the investigators to respond to multiple requests for information. The NMMAPS investigators are to be commended for extending online electronic access beyond the raw data by including the scientific methods used to carry out the analyses. The iHAPSS project illustrates that the capability to share data and methods should prove useful in other carefully considered cases with broad relevance to public health, regulatory, and stakeholder interests. Given costs and other considerations, this approach would not be warranted under all circumstances. Any such venture should only be supported by HEI for the period during which the scientific and stakeholder communities maintain significant interest.

APPENDIX A: HEI Policy on the Provision of Access to Data Underlying HEI-Funded Studies

The provision of access to data underlying studies of the health effects of air pollution is an important element of ensuring credibility, especially when the studies are used in controversial public policy debates. The open and free exchange of data is also an essential part of the scientific process. Therefore, *it is the policy of the Health Effects Institute to provide access expeditiously to data for studies that it has funded and to provide that data in a manner that facilitates review and validation of the work but also protects the confidentiality of any subjects who may have participated in the study and respects the intellectual interests of the investigator in the work.*

This policy applies to all research funded by HEI, whether that research was funded prior to or after November 8, 1999, when amendments to OMB Circular A-110 took effect to require access under the federal Freedom of Information Act (FOIA) to data from federally-supported research that was used in developing a federal agency action that has the force and effect of law.

In responding to FOIA requests through the U.S. EPA or other federal agency, for HEI data that are subject to the Circular A-110 amendments, HEI will follow the principles established in the amendments.

In responding to non-FOIA direct requests to HEI for data, HEI will in general follow the principles described below, which are designed to be consistent with the principles contained in the recent A-110 Amendments, although specific cases may require other arrangements for providing access.

1. *Data* The data to be provided will vary from study to study, but in general will consist of the recorded factual material commonly accepted in the scientific community as necessary to validate research findings. It will not include any of the following: Preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. The "recorded" material excludes physical objects (eg, laboratory samples). Research data also excludes (a) trade secrets, commercial information, materials necessary to be held confidential by a researcher until published, or similar information which is protected under law; and (b) personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study. In some cases, where all of the data used is from publicly available data sets and the

analytic data set can readily and expeditiously be recreated, HEI and/or the Investigator might provide detailed descriptions of how to access and use these public data sets to recreate the analytic data set in lieu of providing the full analytic data set.

2. *Timing* HEI will seek to provide access to data as expeditiously as possible after the completion and publication of the HEI Research Report (or Reports) resulting from the study. In doing so, HEI will, to the maximum practical extent, take into consideration the legitimate intellectual interests of the investigator to have the opportunity to benefit from his or her intellectual endeavors and to publish subsequent analyses from the data set (including additional analyses funded by HEI). In some cases (eg, for studies of particularly high regulatory importance being used to inform decisions over a short time frame), HEI may need to balance the investigator's interests against the need for interested parties to obtain access in a timely manner.
3. *Responsibility and Reimbursement for Costs* To the maximum extent possible, HEI will encourage the Principal Investigator to be the primary sharer of the data. To the extent that providing the data would place an undue burden on the Investigator (eg, in a situation where the sheer number of requests would not allow the Investigator to continue to conduct her or his research), HEI will be prepared to establish an alternative procedure for it to share the data. In either case, HEI will expect to receive from data requesters reasonable reimbursement for both the direct costs of providing the data and for the time of the Investigator and/or HEI staff to gather, transmit, and explicate the data. In order to facilitate data access for all future and current studies in which HEI and the investigator expect that the results have a high likelihood of being used in supporting a regulatory decision, HEI will consider requests from the investigator for a reasonable budget of data archiving funds, to be provided as part of the project budget.
4. *Confidentiality* Any requester of data will be expected to obtain and adhere to all confidentiality approvals necessary to handle the data from the appropriate agencies (eg, the National Center for Health Statistics). HEI will not knowingly provide, or require an investigator to provide, information that can be used to identify a specific individual.
5. *Responsibility of the Data Requester* In addition to the payment of reasonable costs and obtaining any necessary confidentiality approvals, HEI will ask the data requester, as would be normal courtesy in the scientific community, to inform both the Principal Investigator and HEI of any findings emerging from their analysis, to provide the Principal Investigator with an opportunity to respond to those findings prior to publication, to provide copies to both the Principal Investigator and HEI of any papers submitted for publication that used the data, and to cite both HEI and the Principal Investigator in any such publication, noting explicitly that the views expressed are those of the new analyst and not those of the Principal Investigator, HEI, or HEI's sponsors.
6. *HEI Decision Making* All requests for data will be reviewed and decided upon by a Committee of the HEI Science Director and the Chairs of the HEI Research and Review Committees, in consultation with both the research and review staff scientists responsible for the study in question. Any significant policy questions arising from a particular request will be considered by the Board of Directors, upon recommendation of the Committee and the President.

The provision of data will not be simple to accomplish and will at times raise concerns and controversy from one or more parties. HEI will attempt to provide data in a manner that, to the maximum extent practical, fosters an atmosphere of collegiality and mutual respect among all parties, with the aim of obtaining from the sharing of data the maximum benefit for science and for the quality of the public policy decision-making process.

REFERENCES

American Association for the Advancement of Science, Center for Science, Technology, and Congress. 2005. The Shelby Amendment (last updated 02/10/2005). *www.aaas.org/spp/cstc/briefs/accesstodata/index.shtml.* Accessed 05/09/2006.

Daniels MJ, Dominici F, Zeger SL, Samet JM. 2004. The National Morbidity, Mortality, and Air Pollution Study: Part III. PM_{10} concentration–response curves and thresholds for the 20 largest US cities. Research Report 94. Health Effects Institute, Boston MA.

Dominici F. 2004. Time-Series Analysis of Air Pollution and Mortality: A Statistical Review. Research Report 123. Health Effects Institute, Boston MA.

Dominici F, Zanobetti A, Zeger SL, Schwartz J, Samet JM. 2005. The National Morbidity, Mortality, and Air Pollution Study: Part IV. Hierarchical bivariate time-series models: A combined analysis of PM_{10} effects on hospitalization and mortality. Research Report 94. Health Effects Institute, Boston MA.

Health Effects Institute. 1999. Diesel Emissions and Lung Cancer: Epidemiology and Quantitative Risk Assessment. Special Report. Health Effects Institute, Cambridge MA.

Health Effects Institute. 2003. Revised Analyses of Time-Series Studies of Air Pollution and Health. Special Report. Health Effects Institute, Boston MA.

Krewski D, Burnett RT, Goldberg MS, Hoover K, Siemiatycki J, Jarrett M, Abrahamowicz M, White WH. 2000. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. A Special Report of the Institute's Particle Epidemiology Reanalysis Project. Health Effects Institute, Cambridge MA.

National Research Council (US). 1997. Bits of Power: Issues in Global Access to Scientific Data. National Academy Press, Washington DC.

Neutra RR, Cohen A, Fletcher T, Michaels D, Richter ED, Soskolne CL. 2006. Toward guidelines for the ethical reanalysis and reinterpretation of another's research. Epidemiol 17(3):335–338.

Peng, RD, Dominici F, Zeger SL. 2006. Reproducible epidemiologic research. Am J Epidemiol 163(9):783–789.

Samet JM, Zeger SL, Berhane K. 1995. Particulate Air Pollution and Daily Mortality: Replication and Validation of Selected Studies; the Phase I.A Report of the Particle Epidemiology Evaluation Project. Special Report. Health Effects Institute, Cambridge MA.

Samet JM, Zeger SL, Kelsall JE, Xu J, Kalkstein LS. 1997. Particulate Air Pollution and Daily Mortality: Analyses of the Effects of Weather and Multiple Air Pollutants. The Phase I.B Report of the Particle Epidemiology Evaluation Project. Special Report. Health Effects Institute, Cambridge MA.

Samet JM, Dominici F, Zeger SL, Schwartz J, Dockery DW. 2000a. The National Morbidity, Mortality, and Air Pollution Study: Part I. Methods and methodologic issues. Research Report 94. Health Effects Institute, Cambridge MA.

Samet JM, Zeger SL, Dominici F, Curriero F, Coursac I, Dockery DW, Schwartz J, Zanobetti A. 2000b. The National Morbidity, Mortality, and Air Pollution Study: Part II. Morbidity and mortality from air pollution in the United States. Research Report 94. Health Effects Institute, Cambridge MA.

HEALTH
EFFECTS
INSTITUTE

Charlestown Navy Yard
120 Second Avenue
Boston MA 02129-4533 USA
Phone +1-617-886-9330
Fax +1-617-886-9335
www.healtheffects.org

COMMUNICATION 12
October 2006