



UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D.C. 20460

OFFICE OF CONGRESSIONAL AND
INTERGOVERNMENTAL RELATIONS

OCT 30 2013

The Honorable Lamar Smith
Chairman
Committee on Science, Space and Technology
U.S. House of Representatives
Washington, D.C. 20515-6301

Dear Mr. Chairman:

I am writing today to follow up on the commitments we made in our letter of July 30, 2013, to keep you apprised of certain information related to your interest in research data from certain epidemiological studies.

Enclosed, please find: a letter from Harvard University, dated September 25, 2013; a letter from Brigham Young University, dated August 1, 2013; a letter from the American Cancer Society, dated August 19, 2013; a letter from the Health Effects Institute, dated August 27, 2013; and, a letter from the Harvard School of Public Health, dated September 6, 2013.

Please feel free to contact me if you have any questions, or your staff may contact Tom Dickerson in my office at dickerson.tom@epa.gov or (202) 564-3638.

Sincerely,

A handwritten signature in blue ink that reads "Laura Vaught".

Laura Vaught
Associate Administrator

Enclosures

cc: The Honorable Eddie Bernice Johnson
Ranking Member



HARVARD UNIVERSITY
Office for Sponsored Programs

Catherine Breen
Senior Director, Office for Sponsored Programs
catherine_breen@harvard.edu

Holyoke Center, Suite 635
1350 Massachusetts Avenue
Cambridge, MA 02138

t.617.495.9047
f.617.496.2524

September 25, 2013

BY FEDERAL EXPRESS

Mr. Lek Kadeli
Principal Deputy Assistant Administrator
Office of Research and Development
U.S. Environmental Protection Agency
Room 41209
1300 Pennsylvania Ave NW
Washington, DC 20004

Dear Mr. Kadeli:

I am writing on behalf of Harvard University in response to the letter that you sent to Professor Francine Laden on July 8, 2013. Your letter transmitted a request that your agency had received from Senator David Vitter relating to several epidemiological studies on the health effects of certain kinds of air pollution, including a 2006 article written by Prof. Laden and other Harvard researchers ("Reduction in Fine Particulate Air Pollution and Mortality." *American Journal of Respiratory and Critical Care Medicine*. 173: 667-672). According to your letter, the EPA has committed to engaging with Prof. Laden and other researchers to understand what information may be available in response to the Senator's request.

As an institution of higher education focused on teaching, research and scholarship, Harvard believes in and advocates for the exchange of data to advance scientific knowledge. At the same time, we have a responsibility to protect not only individual privacy but also our researchers' intellectual property – interests explicitly recognized in both the Freedom of Information Act and the Shelby Amendment. See 5 U.S.C. § 552(b)(4) and (6); 5 U.S.C. § 552(b)(6).2 C.F.R. § 215.36(d)(2)(i).

Large long-term epidemiological studies, like the air pollution research in question, rely on the participation of thousands of human participants. Without assurances that their private medical and other identifying information will be protected, people would not agree to be part of such studies. In this case, Harvard researchers promised to ensure confidentiality not just to the participants themselves, but also to federal and state agencies.

Moreover, for science to flourish, we must recognize and protect researchers' thought processes, innovative ideas, unique approaches and research designs. Under the Shelby Amendment, for example, research data is defined as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings;" it does not include "preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues." 2 C.F.R. § 215.36(d)(2)(i). Likewise, programs and software that researchers have written are not considered research data.


Your letter recognizes that, in March 2012, after receiving a request pursuant to the Shelby Amendment, Harvard provided to the EPA research data relating to the 2006 article cited by Senator Vitter, and further notes that the EPA subsequently gave a copy of what was provided to Senator Vitter. As required by our confidentiality obligations, this data set did not include individually identifiable information about study participants, nor would Harvard provide such information now.

Moreover, the Krewski report cited in Senator Vitter's request (Krewski, Burnett, et al., 2000. "Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality, Special Report to the Health Effects Institute") (the "HEI Report") itself contains a comprehensive description of the data collection procedures and an audit of the original data from the Harvard Six Cities study, which was the basis for Prof. Laden's 2006 reanalysis. *See generally* HEI Report Part 1: Replication and Validation at 41-130. For example, the HEI Report specifically describes efforts to review original study protocols (at 42 and 94), describes the data processing and quality control (at 42-64), and provides a detailed review of the death certificate coding protocols (at 47-49). A copy of the questionnaire used in the Harvard Six Cities study is reprinted (at 99-114), along with the questionnaire code book (at 115-16). Thus, Senator Vitter already has access to much of the information he is now requesting.

It is also worth noting that a great deal of time has elapsed since data collection began in these long-term air pollution studies. Existing electronic data from the early years of the Harvard Six Cities study may have deteriorated, or may be stored on media that cannot now be read or deciphered by any available devices or software.

I hope this information is helpful to you. If you have any questions or comments, please do not hesitate to contact me.

Sincerely,



Catherine Breen

ECONOMICS DEPARTMENT
130 FACULTY OFFICE BUILDING
BRIGHAM YOUNG UNIVERSITY
PROVO, UT 84602-5535
(801)422-2859



August 1, 2013

Lek Kadeli
Principal Deputy Assistant Administrator
Office of Research and Development
United States Environmental Protection Agency
Washington, D. C. 20460

RE: Requests for data, protocols, methods and related information pertaining to specific epidemiology studies of air pollution and human health.

Dear Lek Kadeli:

I am writing regarding the requests for data, protocols, methods and related information pertaining to specific epidemiology studies on the health effects of particulate matter and ozone air pollution of which I have served as a principle or co-investigator. Details of this request are discussed in your letter dated July 8, 2013 to me and are detailed in the request from Senator Vitter's staff listing the studies and materials requested.

Harvard Six-Cities Cohort Study: Although I was a co-investigator on the initial study of long-term exposure to air pollution and mortality risk (Dockery et al. 1993¹), data analysis was conducted on site while at Harvard. I have not been a co-investigator on the extended follow-up studies of the Harvard Six-Cities cohort (including Laden et al. 2006², Schwartz et al. 2008³, Lepeule et al. 2012⁴) and I do not currently have copies of or direct access to this study's data files. I note, however, that the Kreski et al. 2000 Health Effects Institute (HEI) reanalysis report⁵ and its appendices provide documentation of the Harvard Six-Cities Cohort study that includes an independent data audit, replication of the results of the initial study, copies of the questionnaires and codebook, computer programs and output used in the replication of the original analysis, and related information. The extended follow up studies of the Harvard Six-Cities studies²⁻⁴ provide even further important documentation, replication, and important extensions of the Harvard Six-Cities cohort study.

American Cancer Society Cohort study: The American Cancer Society Cancer Prevention Study II (ACS CPS-II) cohort data were collected by the ACS. The original ACS CPS-II cohort study of long-term exposure to air pollution and mortality (Pope et al. 1995⁶) was a collaborative research effort with ACS researchers. Data analyses occurred on site at the ACS in Atlanta. As

part of the extensive HEI sponsored re-analyses, the ACS made data sharing agreements that allowed separate data access by a large, independent reanalysis team headed by Dr. Dan Krewski at the University of Ottawa to conduct data auditing, replication of originally published results, and substantial sensitivity analyses. For complete documentation, see Krewski et al. 2000.⁶ After the re-analysis report was published in 2000, I collaborated on various research projects with researchers from ACS, University of Ottawa, UC Berkeley and elsewhere that was designed to further extend and document the analysis of the original ACS cohort study and the ACS re-analysis. The ACS CPS-II cohort data used in these studies have remained under the ownership of the ACS. Data analyses has been conducted consistent with maintaining the privacy and confidentiality of research participants and data sharing agreements with ASC. As an external co-investigator collaborating with the ACS, I am not authorized nor am I able to provide any ACS CPS-II cohort data files.

With regards to requests for study protocols, statistical methodologies, questionnaires, and related information pertaining to our studies of air pollution and mortality using the ASC CPS-II cohort, we have and continue to provide substantial documentation in various published and peer-reviewed papers and research reports. Most of the publications are journal articles (including Pope et al. 2002⁷, Pope et al. 2004⁸, Jerrett et al. 2005⁹, Jerrett et al. 2009¹⁰, Turner et al. 2011¹¹, Jerrett et al. 2013¹²) that are necessarily brief (but sometimes include additional documentation in the form of electronic appendices). Others are published as relatively large reports (Krewski et al. 2000⁶, Krewski et al. 2009¹³, Jerrett et al. 2011¹⁴) with even more extensive documentation. Various statistical and other methodological approaches developed for and/or used in these analyses have generally been publically documented in multiple publications and are cited in the journal articles and research reports. Copies of the questionnaires and codebook used in the ACS study are published in the Krewski et al. 2000 HEI report⁶. Available on request to the HEI are appendices that include information regarding computer programs and output used in the replication of the original analysis, the quality assurance audit of the data, occupational exposures, flexible modeling of effects of fine particles and sulfate on mortality, alternate air pollution data, selection of ecologic covariates, definition of metro areas, values of the ecologic covariates, spatial analyses, and random effects Cox models. The questionnaires and other documentation for the ACS cohort are also publically available directly on line. (For a general documentation of the American Cancer Society Cancer Prevention Study II see:

<http://www.cancer.org/research/researchtopreventcancer/currentcancerpreventionstudies/cancer-prevention-study>

For the study questionnaires see:

<http://www.cancer.org/research/researchtopreventcancer/cancer-prevention-questionnaires>).

U.S. Life Expectancy study: The study of reduction in fine particulate air pollution and life expectancy in the U.S. (Pope et al. 2009¹⁵) utilizes data from public sources. The life expectancy data were generated using publically available data as documented in a published paper (Ezzati et al. 2008¹⁶) and the complete data set for the generated life expectancy data is directly available on line at

<http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0050066#s5> .

The socio-demographic and other variables used in the analysis are also directly available from public sources clearly referenced in the paper. For those who do not want to reconstruct the data from original publically available data sources, we have also provided an analytic data file (in

Excel Spread Sheet format and with a complete data dictionary) that includes the full data for the 211 counties in the analysis, that can and has been used to reproduce the paper's results using standard statistical software. These data have been provided under separate cover from Harvard University. Additional published papers have provided extended discussion of methodology and protocol (Pope et al. 2012¹⁷), provided sensitivity analysis regarding potentially influential observations and statistical outliers (Krstic 2012¹⁸, Pope et al. 2013¹⁹) and have provided some expanded and extended analysis (Correia et al. 2013²⁰).

I appreciate the importance of continued efforts to more fully understand the effects of air pollution on human health. I am also fully supportive of open, collaborative, efforts to use data and information in such a way that truly contributes to our scientific understanding, that does not violate the privacy and confidentiality of research participants, that maintains the integrity of the data, and that respects responsible and appropriate sharing of data and replication of results.

Sincerely,



C. Arden Pope III, PhD
Mary Lou Fulton Professor of Economics
Brigham Young University

References:

1. Dockery DW, Pope CA III, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG Jr, Speizer FE. An association between air pollution and mortality in six U.S. cities. *The New England Journal of Medicine* 1993;329:1753-59.
2. Laden F, Schwartz J, Speizer FE, Dockery DW. Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities study. *American Journal of Respiratory and Critical Care Medicine* 2006;173:667-672.
3. Schwartz J, Coull B, Laden F, Ryan L. The effect of dose and timing of dose on the association between airborne particles and survival. *Environmental Health Perspectives* 2008;116:64-69.
4. Lepeule J, Laden F, Dockery DW, et al. Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities Study from 1974 to 2009. *Environmental Health Perspectives* 2012;120:965-970.
5. Krewski D, Burnett RT, Goldberg MS, Hoover K, Siemiatycki J, Jarret M, Abrahamowicz M, White WH. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. Special Report. Health Effects Institute, Cambridge MA, 2000.
6. Pope CA III, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Heath CW Jr. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal of Respiratory and Critical Care Medicine* 1995;151:669-674.
7. Pope CA III, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. Lung cancer, cardiopulmonary mortality and long-term exposure to fine particulate air pollution. *Journal of the American Medical Association* 2002;287:1132-1141.
8. Pope CA III, Burnett RT, Thurston GD, Thun MJ, Calle EE, Krewski D, Godleski JJ. Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. *Circulation* 2004;109:71-77.
9. Jerrett M, Burnett RT, Ma R, Pope CA III, Krewski D, Newbold KB, Thurston G, Shi Y, Finkelstein N, Calle EE, Thun MJ. Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology* 2005;16:727-736.
10. Jerrett M, Burnett RT, Pope CA III, Ito K, Thurston G, Krewski D, Shi YL, Calle E, Thun M. Long-term ozone exposure and mortality. *New England Journal of Medicine* 2009;360:1085-1095.

11. Turner MC, Krewski D, Pope CA III, Chen Y, Gapstur SM, Thun MJ. Long-term ambient fine particulate matter and lung cancer risk in a large cohort of never smokers. *American Journal of Respiratory and Critical Care Medicine* 2011;184:1374-1381.
12. Jerrett M, Burnett RT, Beckerman BS, Turner MC, Krewski D, Thurston G, Martin R, Von Donkelaar A, Hughes E, Shi Y, Gapstur SM, Thun MJ, Pope CA III. Spatial analysis of air pollution and mortality in California *American Journal of Respiratory and Critical Care Medicine* 2013 (in press).
13. Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y, Turner MC, Pope CA III, Thurston G, Calle EE, Thun MJ. Extended follow-up and spatial analysis of the American Cancer Society Study linking particulate air pollution and mortality. HEI Research Report 140, Health Effects Institute, Boston MA. 2009.
14. Jerrett M, Burnette RT, Pope CA III, et al. Spatiotemporal analysis of air pollution and mortality in California based on the American Cancer Society Cohort: Final Report. California Air Resources Board. November 2011.
15. Pope CA III, Ezzati M, Dockery DW. Fine-particulate air pollution and life expectancy in the United States. *New England Journal of Medicine* 2009;360:376-386.
16. Ezzati M, Friedman AB, Kulkarni SC, Murray CJ. The reversal of fortunes: trends in county mortality and cross-county mortality disparities in the United States. *PLoS Med* 2008;5:e66.
17. Pope CA III, Ezzati M, Dockery DW. Validity of observational studies in accountability analyses: the case of air pollution and life expectancy. *Air Quality, Atmosphere & Health* 2012;5:231-235.
18. Krstic G. A reanalysis of fine particulate matter air pollution versus life expectancy in the United States. *Journal of the Air & Waste Management Association* 2012;62:989-991.
19. Pope CA III, Ezzati M, Dockery DW. Fine particulate air pollution and life expectancies in the United States: the role of influential observations. *Journal of the Air & Waste Management Association* 2013;63(2):129-132.
20. Correia AW, Pope CA III, Dockery DW, Wang Y, Ezzati M, Dominici F. Effect of air pollution control on life expectancy in the United States: an analysis of 545 U.S. counties for the period from 2000 to 2007. *Epidemiology* 2013;24(1):23-31.



August 19, 2013

Lek Kadeli
Principal Deputy Assistant Administrator
Office of Research and Development
Environmental Protection Agency
Washington, DC 20460
VIA E-MAIL

Dear Mr. Kadeli:

Thank you for your letter of July 8, inquiring about the permissibility of sharing research data used in certain epidemiological studies focusing on the health effects of particulate matter and ozone pollution. The following is the American Cancer Society's (the Society's) response to your questions.

For 100 years, the Society has worked tirelessly to save lives and create a world without cancer. Along with millions of supporters—over one million of whom volunteered to participate in our research studies—we have committed ourselves to eliminate cancer as a major public health problem. We have been able to lead the way in cancer research by building a foundation of trust with the public and by always placing the public good at the forefront of our mission.

Your inquiry appears to focus on Cancer Prevention Study II (CPS-II) data that were used in four of the studies listed in your letter: Krewski et al (2000),¹ Pope et al (2002),² Jerrett et al (2009),³ and Krewski et al (2009).⁴ CPS-II data were not used in the other studies you identified.

**What Is CPS-II and
Why Are the Data So Valuable?**

The Society established CPS-II in 1982. Over the last 31 years, through the recruitment of nearly 1.2 million male and female participants by approximately 77,000 volunteers in 50 states, the District of Columbia and Puerto Rico, the Society has amassed this data set as a powerful tool to identify the risk factors for cancer and, ultimately, learn how to prevent it. CPS-II data contain comprehensive demographic information as well as health, personal habit history, and economic information. Mortality follow-up of the entire CPS-II cohort continues today with biennial linkage to the National Death Index. The Society has also followed up with subgroups of the larger cohort in a variety of ways, including through repeat questionnaires for assessing cancer incidence and other information and the collection of blood samples and buccal cells for genetic analysis. In addition, Society epidemiologists recently began the retrospective and

stay well | get well | find cures | fight back | cancer.org | 1.800.227.2345

National Home Office
250 Williams Street, Atlanta, GA 30303-1002
404.329.7740 f) 404.329.7530

prospective collection of breast, colorectal, hematopoietic and prostate cancer tumor specimens. In short, the CPS-II data set is one the most comprehensive longitudinal data sets in existence.

CPS-II data and corresponding follow-up studies using the data have played a major role in cancer prevention both nationally and internationally over the past several decades. More than 500 scientific articles have been published and the findings have significantly contributed to our understanding of the health effects of tobacco use, obesity, diet, physical activity, hormone use, and various other exposures in relation to cancer and other diseases.

The value to science and the public of the CPS-II data is incalculable. It is a very large snapshot of human information as it existed and evolved over a period of time, and it continues to be extremely relevant to scientific inquiry. It is a medical treasure built with the commitment of our donors, volunteers, staff, and, most importantly, CPS-II participants.

Responses to EPA's Specific Questions

- 1. Who owns and/or holds the data necessary to replicate the relevant studies and what are the concerns, if any, associated with making such data publicly available?**

A. Control of data

The Society owns, holds and is entrusted with the stewardship of the individual-level CPS-II data. The Society funded and oversaw the collection of the data, and now directs and controls their dissemination. We obtained some of the mortality data in the CPS-II data set from the Centers for Disease Control and Prevention, which manages the nation's National Death Index (NDI). As we explain below, the Society's use and subsequent disclosure of NDI data is limited to those uses and disclosures permitted under NDI's implementing regulations.

The CPS-II data have since been linked, using participant zip codes or other location information, to ecological information about the area in which the subjects lived (the "Linked Analyses"). These Linked Analyses are conducted by Dr. Daniel Krewski at the R. Samuel McLaughlin Centre for Population Health Risk Assessment at the University of Ottawa, under an agreement with the Society to ensure that he and the University handle our individual level data from CPS-II responsibly and ethically.

B. Concerns associated with publicizing data

The Society has a number of serious legal, ethical, and policy concerns regarding disclosure of both the individual level CPS-II data and the Linked Analyses. At the core of our concern is the Society's ethical obligation as steward of personal and highly confidential information. Accordingly, we follow prevailing privacy norms with respect to the data, and we made assurances to participants, the NIH, and the NDI. To provide identifiable data to Congress under these circumstances would violate these legal obligations and commitments. Moreover, the Society's decades-long investment of resources made the collection of CPS-II data possible, and today the data are priceless.

i. *The Society's duty to maintain confidentiality*

a) Certificate of Confidentiality and the National Death Index

The CPS-II data are protected by a Certificate of Confidentiality issued by the NIH to the Society. Under section 301(d) of the Public Health Service Act (42 U.S.C. 241(d)) the Secretary of Health and Human Services may authorize persons engaged in biomedical, behavioral, clinical, or other research to protect the privacy of individuals who are the subjects of that research. This authority has been delegated to the NIH. 42 U.S.C. 241(d). The statute prohibits involuntary disclosure of protected research data:

Persons authorized by the NIH to protect the privacy of research subjects may not be compelled in any Federal, State, or local civil, criminal, administrative, legislative, or other proceedings to identify them by name or other identifying characteristic. 42 U.S.C. 241(d)

If the Society were forced to provide CPS-II data to Congress in direct violation of this statute, the Society would not only breach its Certificate of Confidentiality, but the entire concept of the Certificate and the protection it provides could be in doubt.

Moreover, under these circumstances the Society could not release the information it has received about CPS-II participants' cause of death from the National Death Index, a necessary component of the data to reanalyze the studies in question. The NDI regulations include protections against releasing identifiable information. As we describe in response to Question #2, we are not aware of any way to create a de-identified version of the CPS-II data set sufficient to protect the confidentiality of the participants while at the same time allowing a true replica of the studies.

b) Privacy Policies

The Society is sensitive to and understands the important role of Congress in oversight of environmental policy, but we are concerned that the House of Representatives Committee on Science, Space and Technology's authorization to issue a subpoena for our CPS-II data may put the Society in a position that is inconsistent with prevailing privacy and security standards. Since at least the mid-20th century, confidentiality has been a central tenet of ethical protections for research participants. Individuals share confidential information about themselves to make biomedical and public health research possible and, in exchange, researchers and the public at large assure these volunteers that their confidential data will only be used and disclosed in certain, limited ways. In recent years, these privacy and security protections have become enshrined in various forms, for example in the Health Insurance and Portability and Accountability Act and its implementing regulations, confidentiality protections set forth in the National Death Index regulations, state law, and "privacy by design" principles set forth by the Federal Trade Commission. Although these privacy and security frameworks differ in some respects, core commonalities persist, suggesting a converging set of expectations pertaining to privacy and security.

For example, prevailing privacy norms recognize the need for individuals to be informed about possible permissible uses and disclosures of their data. A closer look at HIPAA is instructive as to legal and public expectations as to privacy. The central tenet of HIPAA is that all uses and disclosures of identifiable data are prohibited, unless they are expressly permitted. Permitted disclosures include those made pursuant to carefully worded authorizations, to *bona fide* researchers under certain, controlled and monitored circumstances, and for public health purposes to health care oversight agencies. HIPAA does not contain any exception to these principles for general congressional curiosity.

Although the Society itself is not directly regulated by HIPAA, most research institutions, such as hospitals and academic medical centers, must comply. The Society is committed to extending the same privacy protections to its research participants as the law would empower institutional providers to extend to their research participants. CPS-II participants deserve no less.

c) Protocols for maintaining confidentiality

Every voluntary participant was assured that their identity and the information they provided, often of a very personal nature, would be kept confidential and used only in connection with research. Volunteers who participated in CPS-II were motivated by a desire to help the fight against cancer and were assured that their commitment and generosity of time and candor would be protected. The confidentiality protections that the Society has in place are vital to the success of research participant recruitment efforts. To balance our promise to the CPS-II participants with our commitment to scientific inquiry, we have a rigorous process to allow outside investigators to request access to CPS-II data subject to confidentiality protections, as explained in our answer to Question #3 below.

ii. *Negative effect on future research*

Violating our legal obligations and breaking the promises we made to participants could damage not only the Society's reputation, but also the next phases of our scientific and public health work. For example, we are currently recruiting participants for our third cancer prevention study ("CPS-3"), and we are concerned that even the threat that Congress might appropriate and possibly make participants' information publicly available could negatively impact our recruitment efforts. More importantly, if research participants believe that confidentiality protections might be limited in circumstances such as these, individuals' willingness to participate in research in all areas may be eroded.

The rationale for the Federal government's acquisition of the CPS-II data appears to be that these underlying data were used in studies that the EPA cited to justify regulatory action. But this sets a dangerous precedent for scientific research: organizations will have reason to fear that any research data cited in connection with a government rulemaking might be subject to confiscation and distribution to the public. This kind of precedent could create a disincentive to researchers to share data, especially if there is a connection to a government rulemaking. Moreover, research entities might limit their own work, choosing to conduct only research that would not be used for government rulemaking to ensure their underlying data are protected. The result could be a breakdown in the collaborative process between scientists necessary to scientific advancement and an impediment to scientific inquiry, particularly in areas of interest for the government. In

addition, this introduces a fundamental disparity in the ethical protections and safeguards for participants in research depending on whether the research is used to inform government policy. What a tragic and ironic disincentive it would be to inform the public that when they give of themselves to support research identified as being of national importance, they must sacrifice basic confidentiality protections.

iii. *Congress cannot properly order EPA to 'take' this data*

The Society's individual level CPS-II data at issue here were funded and collected by the American Cancer Society, and, to the best of our knowledge, without the use of Federal funds. As it is a longitudinal, nationwide study dating from 1982, it is unique and not replicable, and its value cannot be measured. If we were forced by a Committee of the U.S. Congress or by any agency of the executive branch of the federal government to make public this privately created and privately funded resource, it could be akin to taking our property without just compensation in violation of the Fifth Amendment.

iv. *Uncertainty about dissemination caused by Congress is a concern*

Our concerns about confidentiality, the adverse effect on research, and the acquisition of our private property are compounded by statements made about how Congress might disseminate our participants' information. It is our understanding that the House of Representatives Committee on Science, Space and Technology has authorized the Committee Chairman to acquire the CPS-II data by subpoena, if necessary, with the intention of making the data set available "*on the Internet*," as the Chairman stated in an August 1, 2013 public hearing on the subject. The idea that Congress would publish our participants' information online only magnifies our concerns.

- 2. What are the technical options for making these data publicly available, taking into account any concerns about the release of confidential personal health information or other confidential data? What are the implications of these options for replicating these studies? What level of effort in terms of time and resources would be required for these options?**

In order to accurately replicate the studies, Congress will need data and statistical programs that the Society does not hold or control in addition to the raw data in CPS-II. First, Congress will need access to the National Death Index to link the CPS-II data to death records, and to do that, Congress would need the Society to provide participants' name, social security number, date of birth, and state of residence. Then, Congress or others would have to link the appropriate ecological variables to our CPS-II data. Otherwise, Congress will need access to the Linked Analyses, which are maintained by Dr. Daniel Krewski at the University of Ottawa, under an agreement with the Society.

With respect to the Linked Analyses, we do not currently have the internal expertise to determine definitively whether it is possible to code or otherwise modify them in such a way as to protect the confidentiality of our CPS-II participants and also allow for true replication of the studies. To determine what might be possible, we would have to engage outside experts, at considerable expense. This is likely to be a time-consuming and long-term effort with uncertain resolution.

Regarding the CPS-II data, it appears impossible to create a public version that would protect the confidentiality of the CPS-II participants while at the same time allowing a true replica of the studies. To enable study replication, we would have to include individual level information, including participants' location, such as zip code or partial zip code, to enable others to link ecological information. The zip code or partial zip, along with updated zip codes for a portion of the participants, would be listed with a wide variety of personal information, including age, race, gender, education, marital status, height, weight, alcohol consumption, smoking history, exposure to environmental tobacco smoke, occupational history and exposures, and, if applicable, cause of death and death date. Using HIPAA as our guide, we note that zip code *alone* is, in some cases, considered an identifier. Accordingly, the residual zip code information, which is necessary to facilitate the linking with ecological data, combined with other information about each participant, such as race, ethnicity and other data points, would heighten the risk of re-identification. In fact, in light of explosion of publicly available data that can be used to re-identify individuals with data otherwise appearing to be de-identified, regulators continue to expand the single data fields that are classified as "identifiers." While the Society might be able, with sufficient time and resources, to remove all of the confidential or identifying information so that individual CPS-II participants could not be identified, such a data set would be so limited and generic that it would not enable a researcher to replicate the studies in question.

- 3. If there are no feasible options for making all of the data publicly available, how would a researcher gain access to the full set of underlying data in order to replicate these studies? Please provide any documentation you believe would be helpful in understanding this process.**

The Society recognizes the value of externally-proposed studies that are of general interest and high scientific merit. We welcome outside investigators to request access to our data following our application process, the details of which are available on our website.⁵ We only grant access to well-qualified researchers who have demonstrated that their proposed research is well-designed and has the potential to significantly contribute to scientific discourse, and who have the requisite knowledge, qualifications, and experience to conduct the analysis and protect our data.

Once a proposal is accepted, we take various measures to protect our data. Each researcher who is granted access to the data has restrictions on the use and publication of the data and must conduct the research consistent with applicable legal and ethical requirements. Further, a deep understanding of the history of CPS-II and the complexity of the database is needed to conduct scientifically valid research using CPS-II data. Therefore, we require external researchers to work collaboratively with Society investigators, including co-authorship on any resulting publications, and the researchers and their institutions must sign the Society's "Collaboration Agreement," which includes requirements designed to protect the confidentiality of the participants in the research. Moreover, we only give the investigator access to the data that are necessary to conduct the analysis.

The Society may choose to deny requests from individuals sponsored by interest groups who have demonstrated they are not interested in independent and objective scientific research. For example, we have on occasion refused to provide access to scientists who were publicly linked to

sponsorship by tobacco companies. These data are a public trust. We take that responsibility seriously.

We are currently engaged in more than 30 collaborations with outside investigators. With respect specifically to the CPS-II data used for the studies referenced in your letter, I am sure you are aware that the Krewski (2000) study was a replication of original studies precisely because some were concerned about the objectivity related to the results and conclusions of these original studies. As a result of those concerns, the Society shared the necessary data under a confidentiality agreement to ensure our data were properly protected and the reanalysis was done under the auspices of the Health Effects Institute and conducted by a neutral third party.

In summary, the Society has a number of concerns regarding the potential disclosure of our CPS-II data. To compile the CPS-II data set, we assured the 1.2 million individuals who provided personal information to help us understand what causes and prevents cancer that we would maintain the confidentiality of this information. We also applied for and were awarded a National Institutes of Health-issued Certificate of Confidentiality that protects the entire data set, from the date of its inception from disclosure. At the same time, we value the contributions that outside investigators can make using our CPS-II data, which is why we have a process to allow them to apply to access our data subject to confidentiality protocols. Producing CPS-II data to the Federal government outside of our standard process, when we can be given no assurances of how it will be used, by whom, and how widely it would be disseminated, would cause the Society to betray its own policies, the promises it made to participants, covenants with both the NIH and the National Death Index, and prevailing privacy norms. Moreover, the Society has invested countless resources to collect and analyze the CPS-II data, including three decades of work, tens of millions of dollars, and the dedication of 77,000 volunteers. Leaving aside the Society's critical concerns about confidentiality for the citizens who provided personal data, it would be improper for the Federal government to imply appropriate this privately created data set and make it publicly available.

The Society has engaged outside counsel to assist it in protecting the integrity of our CPS-II data. Please include Mr. Stephen M. Ryan of McDermott Will & Emery, LLP and the Society's General Counsel, Mr. Timothy B. Phillips, on all future correspondence. They are the only persons authorized to respond for the Society to any EPA need for further information.

Thank you for your careful consideration of the issues we have raised.

Sincerely,



Otis Brawley, MD, FACP
Chief Medical and Scientific Officer

¹ Krewski D, Burnett RT, Goldberg MS, Hoover K, Siemiatycki J, Jarret M, Abrahamowicz M, White WH. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. Special Report. Health Effects Institute, Cambridge MA, 2000.

² Pope CA III, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. Lung cancer, cardiopulmonary mortality and long-term exposure to fine particulate air pollution. *Journal of the American Medical Association* 2002;287:1132-1141.

³ Jerrett M, Burnett RT, Pope CA III, Ito K, Thurston G, Krewski D, Shi YL, Calle E, Thun M. Long-term ozone exposure and mortality. *New England Journal of Medicine* 2009;360:1085-1095.

⁴ Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y, Turner MC, Pope CA III, Thurston G, Calle EE, Thun MJ. Extended follow-up and spatial analysis of the American Cancer Society Study linking particulate air pollution and mortality. HEI Research Report 140, Health Effects Institute, Boston MA. 2009.

⁵ <http://www.cancer.org/acs/groups/content/@research/documents/document/aCPSc-039148.pdf>



Health Effects Institute

101 Federal Street
Suite 500
Boston MA 02110-1817 USA
+1-617-488-2300
FAX +1-617-488-2335
www.healtheffects.org

August 27, 2013

Mr. Lek Kadeli
Principal Deputy Assistant Administrator
Office of Research and Development
U.S. Environmental Protection Agency
Washington, DC 20460

Dear Mr. Kadeli:

I am pleased to provide you with the response from the Health Effects Institute (HEI) to your letter of July 8, 2013, seeking HEI's advice and comment on the important questions of sharing the data underlying epidemiologic studies of air pollution and health.

As you know, HEI has a longstanding policy to make data underlying its studies available to the widest possible scientific audience. We accomplish this first by the publication of comprehensive, intensively peer-reviewed reports of all results of research we fund (not just those that investigators might select for publication in a peer-reviewed journal), and by making extensive additional details available on-line. We also endeavor, in cases where we have full ownership of and rights to data produced for our studies, to make those data widely available to other investigators, including publishing entire data sets and analytical programs on the web. While there are legitimate privacy concerns that must be addressed in making epidemiologic data with personal health and other information available to other scientific investigators, HEI has long believed that mechanisms can often be developed for doing so and it is the interest of science, and the public policy informed by such science, to find ways to do that.

It is in this spirit that we respond to your letter. We have both several general comments on the nature of the data, and observations on how data may be shared and results replicated, for the particular studies you cite which rely on the American Cancer Society Cancer Prevention Study II and Harvard Six Cities cohorts. We provide, as well, specific answers to your questions.

General Considerations on the Data

As you note in your letter, air pollution epidemiology studies normally rely on several types of data: air quality data, census-based covariate data (e.g. income levels within a zip code area where the study subject(s) reside), health event data (which in these studies are data from the National Death Index), and individual health and personal characteristics data (e.g. level of education, alcohol consumption, body mass index, and smoking behavior) which are gathered through detailed individual questionnaires and in some cases periodic health examinations. We have several general observations:

- Data sets that have been created from publicly available sources and contain no individual identifying information, such as air quality monitoring data and census-based covariate data, should be able to be made publicly available without tremendous difficulty or cost.
- Data from the National Death Index (NDI) – maintained by the Centers for Disease Control and Prevention – is generally made available to investigators upon certification on their part that they would not advertently or inadvertently release the identity or cause of death or any other identifying information of any individual. The NDI does make provisions for making its data available more broadly, but according to well-specified rules for aggregating the data and removing certain information (e.g. specific date of death), which would keep a third party from using the data to identify an individual.
- Data collected from individual subjects in a study which normally includes detailed personal, health status, and behavioral information, is critical to allowing for these studies to determine whether some other factor than air pollution (e.g. obesity or smoking behavior) may be responsible for any health effects that are observed. This data, which is normally collected through individual questionnaires and/or medical examinations, is collected with the *express commitment to the participants - from the organizations and the original investigators that collect the data - that the participants' personal information and identity will not be divulged*. Studies using this data are also subject to the Common Rule, under which investigators must apply to their respective Institutional Review Boards (IRBs) to ensure the protection of human subjects in biomedical and behavioral research.

Observations on Data Sharing and Full Replication of These Studies

The ACS and Harvard studies, at their root, attempt to determine whether persons living in higher pollution areas are more likely to have higher relative risks of premature mortality than those living in lower pollution areas, while attempting to control for a host of personal-level and community-level covariates that may also differ between the individuals and the communities. This by its nature requires knowing where the person lives, which can pose challenges for protecting the identity of an individual if s/he lives in a smaller or sparsely populated area. This challenge has been long recognized, and there are a number of protections in federal rules and scientific practice that address this (e.g. the Census Bureau will not release certain data at the block or even zip code level if they believe that would allow identification).

Since the goal should be to find ways to share data which enables full replication and sensitivity analysis of original studies, it is valuable to consider two aspects of these particular studies that have moved them towards using data at smaller spatial scales:

- First, in response to valid criticisms that the earlier versions of these studies relied only on central air quality monitoring data to estimate exposure, investigators have increasingly sought to better estimate exposure employing land use regression models and other methods that can account for the distance of a subject's home from roadways, industrial facilities, and other sources of air pollution. They have also applied increasingly finer-grained community-level covariates (e.g. at the zip code level). While in the largest locations the application of these finer-grained data would likely not allow

for identification of individual subjects, the national analyses in some of these studies include subjects from a wide range of community sizes, including smaller communities where identification could be possible.

It should be possible to produce a data set which uses techniques like land use regression to assign exposure levels to each subject in a study and to provide only that exposure value in a dataset made available to others. This would avoid the possibility of identification of an individual subject, and would allow for replication of the original results for a study that was analyzing a range of exposure across a specific metropolitan area, for example. But such a data set, absent location information for each participant, would not allow for sensitivity analyses applying different forms of exposure modeling nor full testing of the validity of the original study's exposure estimates.

- Second, as these studies have been reviewed intensively by the HEI Review Committee, the Committee has identified two potentially significant sources of uncertainty in their results: so-called “ecological confounding”¹ and “spatial autocorrelation.”² This is detailed in the HEI Review Committee’s Commentary on the most recent HEI Research Report of Extended Analyses in the American Cancer Society cohort (pp. 128-129 in Krewski 2009). To address both of these issues, one of the first steps that investigators have taken has been to use data at smaller scales, e.g. at the zip code level, which while enhancing their ability to test for these two sources of uncertainties, also poses the potential in smaller communities for individuals and their personal information to be identified.

Taken together, these characteristics – which have in general enhanced the quality and the sensitivity of the studies – increase the difficulty of providing a fully “de-identified” data set while *also* enabling a different investigator to conduct a full replication and sensitivity analysis of the original study results.

Options for Making Data Available – Answers to your Specific Questions

With these considerations in mind, we attempt to answer your specific questions below:

1) Who owns and/or holds the data necessary to replicate the relevant studies and what are the concerns, if any, associated with making such data publicly available?

The publicly available air quality and census covariate data are of course collected and owned by the government and are freely available. The air quality and census data sets created specifically by investigators for a particular study are generally the property of the investigators, but should be capable of being made available, especially in the case where they were created using public funds.

¹ Ecological confounding arises when some community-level variables, which are themselves risk factors for mortality, are also associated with air pollution levels

² Spatial autocorrelation is the tendency for variables to have similar values for people or areas that are geographically close, which can suggest that there are other mortality causes which are unaccounted for in the analysis, or can distort the precision of risk estimates.

As to the ownership of the detailed participant data in the ACS and Harvard Six Cities cohort studies, HEI will leave the answers to the other two recipients of your letter – Harvard University and the American Cancer Society – who created these data sets, maintain them, and would have the most current information on others who may be holding these datasets in whole or in part. Those organizations also provided study participants with express commitments that their personal identity and information would not be divulged and have the responsibility to ensure that this commitment is not compromised during any data sharing.

2) What are the technical options for making these data publicly available, taking into account any concerns about the release of confidential personal health information or other confidential data? What are the implications of these options for replicating these studies? What level of effort in terms of time and resources would be required for these options?

3) If there are no feasible options for making all of the data publicly available, how would a researcher gain access to the full set of underlying data in order to replicate these studies? Please provide any documentation you believe would be helpful in understanding this process.

We see a range of options for making such data available, in different formats and with different procedures, so we are answering the questions jointly. In our view, it is feasible to share data in one of three ways (which have been used in many instances) and to do so while protecting the privacy of the individual subjects. The options range, however, from those that offer the most detailed access to study data to those that offer significantly less access:

A. Collaboration with original investigators to obtain full access to data in order to conduct joint analyses

This process is the most common practice in the scientific community for sharing personal data. It normally involves either formal or informal application processes for a scientific researcher to ask the original organizations and investigators who created the data set to gain access to the data to allow for collaborative analyses of an important research question. The American Cancer Society, for example, provides explicit instructions on their website on how to collaborate with them, and many other investigators have conducted more informal collaborations of a similar type. Such collaborations have, of course, to be conducted in full compliance with the Common Rule and any federal or other requirements for protecting the privacy of the participants.

The *advantage* of this process is that it can provide investigators with the fullest access to the data sets and with the benefits of regular consultation with the original investigators whenever there are questions about data structure or content. The *disadvantages* include that the original investigators may not choose to collaborate with all who request access, and a fully independent replication and sensitivity analysis of the original studies may not be possible or broadly accepted, given the collaborative relationship.

B. Application to obtain independent access to analytic data sets sufficient to allow for replication and sensitivity analysis of the original results

This process involves the request by a researcher to the original investigators, or to agencies and organizations, who created the data set to gain access to the data sets underlying a particular study. This normally would involve the development of a protocol for such analysis by the researcher, the review and approval of the protocol by the submitting scientists' IRB, explicit signed commitments by the researchers that they will not disclose personal information (on pain of penalty in the case of federally owned data sets), and usually other protections (e.g. prohibition of the publication of any results presenting data for groups of fewer than a certain number of subjects, and review by the original investigators before publication to ensure that no such information is inadvertently disclosed). Such a process is currently used within the US Department of Health and Human Services.

One relevant example of such data sharing is the detailed data sharing procedures established for the Multi-Ethnic Study of Atherosclerosis (MESA) which can be viewed at https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?view_pdf&stacc=phs000403.v1.p3. In addition, MESA has created several "Limited Access Data Sets" in which personal identifying information has been removed and which can be accessed more readily, but which would not allow for full replication of original studies (see <https://biolincc.nhlbi.nih.gov/studies/mesa/?q=MESA>).

The *advantage* to this approach is that it can provide access to a substantial portion of the relevant data and allow for fully independent replication and sensitivity analyses of the original results. The major *disadvantage* is that this approach normally does not provide access to the full data set, but rather only to the detailed analytic data set or summary tables used in specific studies, thus precluding full replication.

A similar albeit much more intensive process enabled HEI and its independent investigators to gain access to the full data which we reanalyzed from the Harvard Six Cities Study and the American Cancer Society Study (HEI 2000). This process was structured to allow intensive efforts to replicate and test the robustness and sensitivity of the originally reported results. It was undertaken with the full agreement of, but not collaboration with, the original investigators, and provided full access to the data in accordance with a specifically developed data use agreement which ensured protection of privacy. The analyses were also informed by expert advisors from industry, academia, and other stakeholders.

C. Provision of a "de-identified" disk (or other electronic medium) to provide a more limited data set that would not under any circumstances allow for identification of individuals

In some cases, the simplest mechanism for providing access to study data would be through the provision of a fully de-identified data set in electronic form that can be readily shared with all parties without the possibility of an individual and his or her personal characteristics to be divulged. This has the *advantage* that it may allow independent replication and sensitivity analyses of some of the results of the original investigators. The most significant *disadvantage* is that, as noted above, the most recent analyses in the ACS populations have applied increasingly finer-grained community level data analysis; the release of a fully "de-identified" dataset will not allow full replication and sensitivity analysis of these most recent results, e.g. the testing of

alternative models for estimating exposure among the study subjects, and the inability to test whether ecological confounding and spatial autocorrelation could be affecting the results.

Overall, HEI believes that the opportunity for other scientific investigators to have access to and conduct additional analyses in these epidemiologic data sets is of tremendous scientific value, and can provide additional understanding of important scientific questions that can in turn inform air quality policy decisions. As we have described, there are well-established processes for making such data available; however, not all processes provide the fullest access to the data required while still protecting the privacy of individual information that is essential to the studies.

We would be pleased to provide additional consultation on these important questions and to answer any questions you might have. Please let us know if you have further questions or need additional assistance in this effort. You may feel free to contact me or HEI Science Director Dr. Rashid Shaikh at rshaikh@healtheffects.org or (617) 488-2301 for any follow-up questions

Sincerely,



Daniel S. Greenbaum
President

cc: Dr. Rashid Shaikh
Dr. Susan Gapstur, American Cancer Society
Dr. Douglas Dockery, Harvard University

Health Effects Institute. 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.

Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y, Turner MC, Pope CA III, Thurston G, Calle EE, Thun MJ. 2009. Extended Follow-Up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality. HEI Research Report 140. Health Effects Institute, Boston, MA.



HARVARD SCHOOL OF PUBLIC HEALTH

Department of Environmental Health
665 Huntington Avenue
Boston, MA 02115-6021

Douglas W. Dockery
Professor of Environmental Epidemiology
Department Chair

Tel. 617 432-0729
Fax 617 277-2382
ddockery@hsph.harvard.edu

September 6, 2013

Mr. Lek Kadeli
Principal Deputy Assistant Administrator
Office of Research and Development
U.S. Environmental Protection Agency
Washington, DC 20460

Dear Mr. Kadeli:

I am pleased to respond to your letter of July 8, 2013, seeking advice and comment on sharing the data underlying epidemiologic studies of air pollution and health. Let me address each of your three questions specifically.

1. Who owns and/or holds the data necessary to replicate studies and what are concerns, if any, associated with making such data publicly available?

This question makes several assumptions which must first be clarified.

WHAT IS MEANT BY REPLICATION?

Replication is the standard for scientific investigations. Replication implies independent data, analytic methods, laboratories, and methods (Greenbaum, Bachmann et al. 2001; Peng, Dominici et al. 2006). While replication is the standard in physical and biological (experimental) sciences, replication can be difficult in epidemiology (observational) sciences where it may be hard to find comparable data from independent populations.

True replication of long-term observational (epidemiologic) studies is time-consuming and costly. Nevertheless, there has been replication of the original air pollution mortality associations reported in the Six Cities study reported in 1993 (Dockery et al., 1993). Indeed, the subsequent analysis of air pollution associations in the American Cancer Society CPS II cohort (Pope et al., 1995) was undertaken explicitly as an independent replication of the observations in the Six Cities study of mortality associations with fine and sulfate particulate matter air pollution (Greenbaum, Bachmann et al. 2001). Since these original observational studies two decades ago, there have been numerous reported replications of the original findings in independent studies from the United States and Europe (see Table below abstracted from a recent review of these studies (Hoek, Krishnan et al. 2013)). The EPA particulate national ambient air quality standard is based on a review of all of this body of evidence and not solely on the Six Cities and ACS studies. The EPA benefit analyses used exposure response functions from these two studies because they represent the range of exposure response reported in the scientific literature.

Reproducing results implies that independent investigators subject the original dataset to their own analyses and interpretation (Peng, Dominici et al. 2006). Reproducing results does not provide the same level of independent replication, but at times can be the only feasible approach.

Several authors have advocated that data and analytic code should be routinely made publically available for epidemiology studies to allow for reproduction of published results (Peng, Dominici et al. 2006; Hernan and Wilcox 2009; Samet 2009).

In 1997, following calls for release of original data for the Six Cities and ACS analyses (Greenbaum, Bachmann et al. 2001), the Harvard and ACS investigators agreed to provide a copy of the analytic datasets and access to the original records to independent investigators selected by the Health Effects Institute, with appropriate assurances and oversight to ensure protection of participants' confidentiality. These data were subjected to validation of the data records, an attempt to reproduce the original results by independent analyses, and testing the sensitivity of the original published results to alternative assumptions, methods, and adjustment for additional potential confounders. This quality assurance check and reanalysis found the data to be of high quality, the results to be reproducible, and the findings to be insensitive to alternative analytic approaches and control of confounders. These results were published in a 293 page peer-reviewed HEI report (Krewski, Burnett et al. 2000), and published in the peer-reviewed scientific literature (Krewski, Burnett et al. 2003; Krewski, Burnett et al. 2005; Krewski, Burnett et al. 2005).

TABLE 1: Long-term cohort studies of the effects of particulate air pollution (PM_{2.5}, PM₁₀, and TSP) on mortality. Abstracted from Hoek, Krishnan et al (2013).

Study	Study population	Follow-up period	Pollutant	Authors	Publication Year
Harvard Six Cities	8111 adults in six US cities	1976 - 1989	PM _{2.5}	Dockery et al	1993
American Cancer Society (ACS) Study	552,800 adults from 51 US cities	1982 - 1989	PM _{2.5}	Pope et al	1995
ACS Study	500,000 adults from 51 US cities	1982 -1998	PM _{2.5}	Pope et al	2002
ACS Sub-Cohort Study	22,905 subjects in Los Angeles area	1982 - 2000	PM _{2.5}	Jerrett et al	2005
Harvard Six Cities	3096 adults in six US cities	1979 -1998	PM _{2.5}	Laden et al	2006
German Cohort	4752 women in Ruhr area	1985 - 2003	PM ₁₀	Gehring et al	2006
Women's Health Initiative Observational Study	65,893 postmenopausal women from 36 US metropolitan areas	1994-1998	PM _{2.5}	Miller et al	2007
Netherlands Cohort Study	120,852 subjects from Netherlands	1987 -1996	PM _{2.5}	Beelen et al	2008
Nurses' Health Study	66,250 women from the US north eastern metropolitan areas	1992-2002	PM ₁₀	Puett et al	2008

Medicare National Cohort	13.2 million elderly Medicare recipients across the USA	2000 - 2005	PM _{2.5}	Zeeger et al	2008
Nurses' Health Study	66,250 women from the US north eastern metropolitan areas	1992-2002	PM _{2.5}	Puett et al	2009
Swiss National Cohort	National census data linked with mortality	2000 - 2005	PM ₁₀	Huss et al	2010
California Teachers Study	45,000 female teachers	2002 -2007	PM _{2.5}	Ostro et al	2010
US Trucking Industry Cohort	53,814 men in the US trucking industry	1985 -2000	PM _{2.5}	Hart et al	2011
Health Professionals Follow-Up Study	17,545 highly educated men in the midwestern and northeastern US	1989 - 2003	PM _{2.5}	Puett et al	2011
China National Hypertension Survey	70,497 men and women	1991 - 2000	TSP	Cao et al	2011
California Teachers Study	101,784 female teachers	1997- 2005	PM _{2.5}	Lipsett et al	2011
Chinese Retrospective Cohort Study	9,941 adults from five districts of Shenyang city	1998 -2009	PM ₁₀	Zhang et al	2011
Vancouver Cohort	452,735 Vancouver residents 45-85 yr	1999 - 2002	PM _{2.5}	Gan et al	2011
Harvard Six Cities	8096 adults in six US cities	1974 - 2009	PM _{2.5}	Lepeule et al	2012
Nippon Data Cohort	7,250 adults > 30 yr throughout Japan	1980 - 2004	PM ₁₀	Ueda et al	2012
Canadian National Cohort	2.1 million nonimmigrant Canadians . > 25 yr	1991 - 2001	PM _{2.5}	Crouse et al	2012
New Zealand Census Mortality Study	1.06 million adults in urban areas from 1996 census	1996 -1999	PM ₁₀	Hales et al	2012
German Cohort	4752 women in Ruhr and surrounding area	1985 - 2008	PM ₁₀	Heinrich et al	2013
Rome Longitudinal Study	1,265,058 adults from Rome	2001 - 2010	PM _{2.5}	Cesaroni et al	2013

WHO OWNS AND/OR HOLDS THE DATA?

Under the terms of the NIEHS grants and EPA contracts, the Six Cities data are owned and held by the President and Fellows of Harvard College. This ownership of the data by Harvard is well established legally.

WHAT ARE CONCERNS WITH MAKING SUCH DATA PUBLICALLY AVAILABLE?

Harvard has supported free exchange of data for reproducing and advancing scientific knowledge whenever individual privacy is not compromised.

A recent example was the release of lung function measurements of children in the Six Cities study collected between 1974 and 1989, for a multinational pooled analysis of normal values for children (Quanjer, Hall et al. 2012; Quanjer, Stanojevic et al. 2012). In this case, individual data including sex, race/ethnicity, age, height, weight, and lung function were released. Individual identifiers were not included and the characteristics released were not alone sufficient to allow identification of individual children.

In asking potential subjects to participate, we assured all participants that their individual data would not be released to anyone other than the study investigators (see below).

In the case of mortality records, there are a variety of standards. Individual death records are compiled by each state, and forwarded to the National Center for Health Statistics (NCHS). Death records are made available to researchers in several forms. Surveillance data of deaths have previously been available by county and death date from the National Center for Health Statistics. While these data sets did not include individual identifiers prior to 1989, they did include sex, age, race/ethnicity, date of death, county of death, and primary cause of death. However, concerns with privacy of death data have led to increasing restrictions on the identifiable data (Centers for Disease Control and Prevention 2013).

Over the years, confidentiality standards have changed for the public release of geographic and date details on vital statistics micro-data files (Centers for Disease Control and Prevention 2013). These changes are reflected in the data available in successive time periods, as follows:

- Prior to 1989, NCHS public-use death micro-data files contained all counties and exact dates (year, month, and date) of deaths.
- Between 1989 and 2004, public-use death micro-data files contained only geographic identifiers of counties and cities with a population of 100,000 or greater, and no exact dates of death (year, month, and day of week, e.g. Monday, only).
- Beginning in 2005, public-use death micro-data files contained individual-level vital event data at the national level only, that is, with no geographic identifiers (no state, county, or city identifiers), and no exact dates of death (year, month, and day of week, e.g. Monday, only).

Thus, since the study was published in 1993 there has been a substantial shift in the standards for confidentiality of death records, as reflected by the practices of the National Center for Health Statistics of the Centers for Disease Control and Prevention.

Since 1979, individual death records have been compiled into the National Death Index, a national resource for follow-up studies. Investigators may apply to the NDI to search for deaths of study participants. NDI requires informed consent of the study participants, institutional review board oversight, and assurances that identifiable data are not released. Standards for release of death data vary between states. In some states, death records are considered public and are readily available. In other states, death records are considered private, and are available only to next of kin (immediate family).

Prior to the creation of the National Death Index, the Six Cities Study investigators had to apply to each state to obtain copies of death certificates. Cause of death was coded by a certified nosologist from the original death certificate. Release of death data was then dictated by the most restrictive state privacy requirements.

EXAMPLES OF REPRODUCTION

In the case of non-identifiable mortality data, Harvard investigators have worked with interested independent investigators to replicate published findings. For example, the 1996 study entitled "*Is daily mortality specifically associated with fine particulate air pollution*" examined the effect of acute air pollution exposures on counts of daily mortality in the Six Cities Study communities (Schwartz, Dockery et al. 1996). In a replication/reanalysis exercise sponsored by the Electric Power Research Institute, independent investigators at Klemm Associates were provided with copies of the original data. They attempted to reproduce the original mortality data, replicate the original analyses, and assess the sensitivity of the analyses to alternative methods and control of covariates. This led to joint (Klemm, Mason et al. 2000) and independent (Klemm, Mason et al. 2000) peer-reviewed publications.

A more recent study examined the association of changes in county-specific life-expectancy with changes in fine particle air pollution in 211 counties in the United States between 1980 and 2000 (Pope et al., 2009). These data were compiled from publically available datasets and included no individual death records. Copies of these data were provided to interested individual investigators including Dr. Goran Krstić of Fraser Health in British Columbia, Dr. James Enstrom of the Scientific Integrity Institute, and Dr. Stanley Young of the National Institute for Statistical Sciences (a private, nonprofit organization in Research Triangle Park, NC). These re-analyses have led to a lively debate in scientific literature. Dr. Krstić published a critique in 2012 (Krstic 2012). Dr. Enstrom presented his reanalysis at a symposium (Enstrom 2010). Dr. Young has presented his results orally (Young 2010) and more recently in the peer-reviewed literature (Young and Xia 2013). The original authors published responses to these critiques in peer-reviewed journals (Pope, Ezzati et al. 2013), as is normal practice in scientific debate.

As these re-analyses illustrate, there has not been a question of availability of mortality/air pollution data when individual death records are not involved.

2. What are the technical options for making these data publicly available, taking into account any concerns about release of confidential personal health information or confidential data? What are the implications of these options for replicating these results? What level of effort in terms of time and resources would be required for these options?

Release of identifiable individual data would violate the assurances of confidentiality required by the Harvard Human Studies Committee (Institutional Review Board) and given to each study participant upon their enrollment into the Six Cities Study. As participants were enrolled into the study, they signed the following "*Assurance of Confidentiality*," also signed by Benjamin G. Ferris, Jr., the Principal Investigator of the study, and by a witness:

Harvard University School of Public Health hereby gives the assurance that your identity and your relationship to any information obtained by reason of your participation in this study of respiratory symptoms will be kept confidential and will not otherwise be

disclosed except as specifically authorized by you. The data from individuals will be pooled and used as group data in scientific studies.

As custodians of these data, we consider that we are obligated to maintain the commitment to maintain this Assurance of Confidentiality made with each participant in the study.

In addition, release of identifiable individual death records would violate the agreements with the National Death Index and with the individual state agencies to obtain copies of the individual death records. For example, the original application requesting data from the National Death Index includes the following *Applicant Assurance*:

The identifiable data obtained from the National Death Index will be used only for research and statistical purposes. With the exception of requests for death record information made to the appropriate State vital statistics office, no data will be published or released in any form if a particular individual or establishment supplying the information or described in it is identifiable.

In addition, we had to apply to each state vital statistics division to obtain copies of death records. In each case, we had to provide assurances of confidentiality of these vital records. For example, the Missouri Division of Health required:

The request will be approved only if adequate assurances are provided to protect the confidentiality of the records requested. This includes limiting access to the records only to members of the research staff, not releasing records to other agencies, publishing data so individuals cannot be identified, destroying the records upon completion of the study, and not contacting family members or acquaintances of decedents or infants without written permission from the Director of the Missouri Division of Health.

Thus we also have made very explicit institutional commitments to protect the confidentiality of the death information of participants in the study.

DATA REQUIRED FOR REPRODUCING RESULTS

What data are required to reproduce the results of the 1993 mortality analyses (Dockery, Pope et al. 1993), the 2006 mortality follow-up (Laden, Schwartz et al. 2006), or the most recent mortality follow-up (Lepeule, Laden et al. 2012)? There are three classes of data required for these analyses: exposures, health outcomes, and the covariates (or confounders). Let us consider each of these separately starting with exposures.

For these analyses, the exposures are community level air pollution concentrations. Air pollution concentrations are publically available. This study included annual mean air pollution concentrations collected specifically for this study at a centrally located site in each community. There is no issue with making these air pollution data publically available. However, to conduct the analysis, the residency of each research participant must be linked to the exposure data, resulting in the identification of the subjects' city of residency.

The health outcome is time to death (or cause-specific death) from the start of the study for each individual. This requires knowing when a person was enrolled in the study, when they died and cause of death, and if they did not die or were lost to follow-up, the date of last contact.

The covariates that need to be considered for reproducing the results are other predictors of death. In this analysis, the covariates included age, sex, race, smoking (indicators of current

and former smoking, number of pack-years smoked), education (indicator of less than high school), and body-mass-index. Defining exposure required knowing city of residence at enrollment into the study. Knowing their individual characteristics alone would not be sufficient to identify an individual in the study. As noted above, these types of non-identifiable data have been released to other researchers. The difficulty arises when these individual characteristics (covariates) are combined with death records (date of death) and exposure information (place of residence).

De-identification is not simply the process of removing names and addresses. To illustrate the difficulty of ensuring privacy with respect to death records, consider a study participant in Watertown, Massachusetts, the first city enrolled in the study. The 1990 census population of Watertown was 33,284. Assuming a national average death rate of 799.5/100,000 per year (Centers for Disease Control and Prevention 2013), we would expect less than one (0.73) death per day. Knowing a participant from Watertown died on a specific date would almost certainly allow identification of that individual from published obituaries, and hence is considered identifiable information. Knowing the person's age, sex, and race as required to reproduce the analyses would leave no doubt of their identity. The table below presents the 1990 census population for each of the Six Cities and estimated numbers of deaths per day.

TABLE 2: 1990 census population in each of communities in the Harvard Six Cities Study, and expected number of deaths per day based on US average death rates (Centers for Disease Control and Prevention 2013)

Study Community	1990 Population	Expected Deaths/Day [†]
Portage/Pardeeville/Wyocena, WI	10,890	0.24
Kingston/Harriman, TN	11,671	0.26
Steubenville, OH	22,125	0.48
Watertown, MA	33,284	0.73
Topeka, KS	119,883	2.63
St. Louis, MO*	396,685	8.69

[†] Assuming US average of 799.5 deaths/100,000/year

*Note: St. Louis sample only included residents of the Carondelet section of St. Louis. Census is for entire city.

Thus knowing the date of death plus the essential individual characteristics for these analyses – sex, age, and city of residence, is sufficient to identify individual study participants. Furthermore, even knowing the year of death, in combination with sex, age and city of residence would be sufficient to identify most participants.

For comparison, as noted earlier, prior to 1989 the National Center of Health Statistics only released public use data specifying date of death and county of residence. This was subsequently changed to specify only counties with population greater than 100,000, and date was reported only as year, month, and day of the week. Currently, public-use death data are only

available without specification of county of residence and no exact dates of death are provided (year, month, and day of week only).

3. If there are no feasible options for making all the data publically available, how would a researcher gain access to the full set of underlying data in order to replicate these studies? Please provide any documentation you believe would be helpful in understanding this process.

First, we would like to note that as indicated above, the results of the Six City Study have been both replicated and reproduced. More broadly, we have struggled with the competing demands of providing full access to policy-relevant observational public health data while maintaining the confidentiality of personal data for more than 15 years. As illustrated in the previous sections, these issues have been the subject of vigorous debate. Based on this experience, we would suggest that there are two approaches to allow independent researchers to gain access to the full set of underlying data.

The first approach would to provide access to all the data as we did in response to the EPA request in 1997. On January 31, 1997, Mary Nichols, EPA Assistant Administrator for Air and radiation wrote to Dr. Dockery stating in part:

"As you know, there has been considerable interest in your research on the health effects of air pollution, including requests by members of Congress, governors of several states, and other for the raw data underlying your published research. ... (G)iven the strong interest in your research, EPA would encourage reasonable accommodations with the scientific and governmental community that would permit other interested scientists and agencies to understand fully the basis for your work. We therefor request that you make data associated with your published studies available to interested parties as rapidly as possible."

After thoughtful consideration of this request, in April 1997 we asked an outside, independent agency, the Health Effects Institute (HEI), to provide an independent, comprehensive review and re-evaluation of the study data. We agreed to turn over a complete copy of all the data and provide access to all original records to HEI. There were no constraints on analyses or questions that the HEI investigators could explore. However, the HEI investigators were required to apply for and receive approval from the same agencies and institutional review boards that approved the original Harvard study that generated these data. In addition, the data were kept on a secure computer, not connected to the web or network, to ensure data security.

HEI assembled an Expert Panel to provide scientific oversight of the reanalysis project. The HEI Expert Panel had an open competition for a team of investigators to conduct the reanalyses. Harvard had no input into the process of selecting the independent scientific review team. A team from the University of Ottawa was selected.

HEI also established an Advisory Board to provide stakeholder participation (Health Effects Institute 2000). HEI solicited and compiled questions broadly through open solicitation and public meetings.

In 2000 the independent investigators produced a report which was peer-reviewed and then reviewed by the Expert Panel. The Harvard investigators were given an opportunity to comment on the report but not to edit it. The report, Expert Panel review, and original investigator comments then were published by HEI (Krewski, Burnett et al. 2000). In addition,

the results have been published in the peer-reviewed scientific literature (Krewski, Burnett et al. 2003; Krewski, Burnett et al. 2005).

This complete access approach provided a transparent review of the quality of the data, reproduction of the original results, and analyses of the sensitivity of the findings to alternative methods and control for alternative explanations. While this process was comprehensive and successful, it was also long and expensive, making it less than an ideal model (Greenbaum, Bachmann et al. 2001). Moreover, since the data integrity and findings of the Six Cities study already has been reproduced, the argument for repeating this process seems weak.

The alternative approach is to allow specific, restricted access to interested investigators. As a groundbreaking study and as a valuable data resource, the Six Cities Study remains a potential resource for additional analyses. The Harvard investigators have been and continue to be open to collaborating with interested, qualified investigators to fully explore the use of these observational data for discovery and better understanding.

Interested investigators may apply to use specific data to address specific questions. This approach has been used in several similar large observational studies.

For example, the American Cancer Society (ACS) has a well-defined procedure for outside investigators to propose questions that could be addressed using the Cancer Prevention Studies (American Cancer Society 2013). Similarly, at Harvard, the Nurses' Health Studies have established procedures for proposing use of the data sets (Nurses' Health Study 2013).

Following the model of the procedures for the American Cancer Society and the Nurses' Health studies, we could create a formal procedure for requesting and monitoring access to data from the Six Cities Study, managing and monitoring analyses, and monitoring dissemination of results.

The first step would be to establish an independent expert panel to establish procedures, review applications, and monitor the process. One option would be to ask the existing *External Advisory Committee* of the Harvard Clean Air Research Center to take on this task.

Requests for access to data would require a formal application to the *External Advisory Committee*. Following the examples of the ACS and Nurses' Health studies, such an application could include the following elements:

- Specific hypothesis of the proposed analysis
- Scientific significance of the project
- Data variables required and analysis plan
- Reasons for proposing use of these data, rather than another source
- Sources of funding
- Qualifications of external investigator
- Identification and agreement of collaborating Harvard investigator

Upon approval of the *External Advisory Committee*, the external and Harvard investigators would enter into a formal agreement, which, again based on ACS and Nurses' studies examples, could include the following elements:

- All primary data, computer programs, and analysis results would be maintained on the Harvard computer servers, and all data analyses will be conducted on Harvard computers.

- Agreement on the role of Harvard collaborator(s) on the project, and authorship for specific publications arising from the work using the Harvard data.
- At least one member of the Harvard investigative team would be a coauthor on any manuscript resulting from this collaboration and, as such, would need to approve any manuscript prior to its submission for publication.
- Certification of Human Subject training for each investigator and approval from the Harvard School of Public Health Human Subjects Committee (Institutional Review Board).
- Prohibited use of the material for any purpose other than that explicitly stated in the proposal.
- Guarantee of the confidentiality of any data arising from the study, and agreement not to release data to any other person or group for any purpose, except with the explicit permission of Harvard investigators.
- Specification of terms for payment for time and effort by Harvard investigators.

As noted above, these procedures have been commonly applied in providing access of interested investigators to similar population based studies, while protecting confidential individual information. Given others' successful experience with this approach, Harvard stands ready to work on such a process with interested investigators.

I hope you find these comments helpful, and I would be pleased to provide additional consultation on these important questions. Please let us know if I can be of further assistance in this effort.

Best regards,



Douglas W. Dockery, ScD

xc: Michael Grusby, Catherine Breen

References Cited

- American Cancer Society (2013). Epidemiology Research Program. Atlanta, GA, American Cancer Society. 2013.
- Centers for Disease Control and Prevention (2013). FastStats: Death and Mortality. Hyattsville, MD, national Center for Health Statistics. 2013.
- Centers for Disease Control and Prevention (2013). NCHS Data Release and Access Policy for Micro-data and Compressed Vital Statistics Files, Division of Vital Statistics, National Center for Health Statistics. 2013.
- Dockery, D. W., C. A. Pope, 3rd, et al. (1993). "An association between air pollution and mortality in six U.S. cities." N Engl J Med 329(24): 1753-9.
- Enstrom, J. E. (2010). Critique of CARB diesel science, 1998–2010. CARB symposium: estimating premature deaths from long-term exposure to PM2.5. Sacramento, CA.
- Greenbaum, D. S., J. D. Bachmann, et al. (2001). "Particulate air pollution standards and morbidity and mortality: case study." Am J Epidemiol 154(12 Suppl): S78-90.
- Health Effects Institute (2000). Preface: Particle Epidemiology Reanalysis Project. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. Cambridge, MA, Health Effects Institute: 1-6.
- Hernan, M. A. and A. J. Wilcox (2009). "Epidemiology, data sharing, and the challenge of scientific replication." Epidemiology 20(2): 167-8.
- Hoek, G., R. M. Krishnan, et al. (2013). "Long-term air pollution exposure and cardio-respiratory mortality: a review." Environ Health 12(1): 43.
- Klemm, R., R. Mason, et al. (2000). "The effect of imputation of exposure estimates on the association between fine particulate matter and mortality." Ann Epidemiol 10(7): 477-478.
- Klemm, R. J., R. M. Mason, Jr., et al. (2000). "Is daily mortality associated specifically with fine particles? Data reconstruction and replication of analyses." J Air Waste Manag Assoc 50(7): 1215-22.
- Krewski, D., R. T. Burnett, et al. (2005). "Reanalysis of the Harvard Six Cities Study, part II: sensitivity analysis." Inhal Toxicol 17(7-8): 343-53.
- Krewski, D., R. T. Burnett, et al. (2005). "Reanalysis of the Harvard Six Cities Study, part I: validation and replication." Inhal Toxicol 17(7-8): 335-42.
- Krewski, D., R. T. Burnett, et al. (2003). "Overview of the reanalysis of the Harvard Six Cities Study and American Cancer Society Study of Particulate Air Pollution and Mortality." J Toxicol Environ Health A 66(16-19): 1507-51.
- Krewski, D., R. T. Burnett, et al. (2000). Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. Special Report. Cambridge, MA: 293.
- Krstic, G. (2012). "A reanalysis of fine particulate matter air pollution versus life expectancy in the United States." J Air Waste Manag Assoc 62(9): 989-91.
- Laden, F., J. Schwartz, et al. (2006). "Reduction in fine particulate air pollution and mortality: Extended follow-up of the Harvard Six Cities study." Am J Respir Crit Care Med 173(6): 667-72.

- Lepeule, J., F. Laden, et al. (2012). "Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009." Environ Health Perspect 120(7): 965-70.
- Nurses' Health Study (2013). Guidelines for External Collaborators: Use of the Nurses' Health Studies Archived Data. Boston, MA, Channing Laboratory, Brigham and Women's Hospital.
- Peng, R. D., F. Dominici, et al. (2006). "Reproducible epidemiologic research." Am J Epidemiol 163(9): 783-9.
- Pope, C. A., 3rd, M. Ezzati, et al. (2013). "Fine particulate air pollution and life expectancies in the United States: the role of influential observations." J Air Waste Manag Assoc 63(2): 129-32.
- Quanjer, P. H., G. L. Hall, et al. (2012). "Age- and height-based prediction bias in spirometry reference equations." Eur Respir J 40(1): 190-7.
- Quanjer, P. H., S. Stanojevic, et al. (2012). "Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations." Eur Respir J 40(6): 1324-43.
- Samet, J. M. (2009). "Data: to share or not to share?" Epidemiology 20(2): 172-4.
- Schwartz, J., D. W. Dockery, et al. (1996). "Is daily mortality associated specifically with fine particles?" J Air Waste Manag Assoc 46(10): 927-39.
- Young, S. (2010). Health findings and false discoveries: Voodoo statistics and trust me science. The American Association for the Advancement of Science (AAAS) 2010 Annual Meeting. San Diego, CA.
- Young, S. and J. Q. Xia (2013). "Assessing geographic heterogeneity and variable importance in an air pollution data set." Statistical Analysis and Data Mining 6: 375-386.