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.