HEI

Health Effects Institute

75 Federal Street Suite 1400
Boston MA 02110-1940 USA
+1-617-488-2300
FAX +1-617-488-2335
www.healtheffects.org

August 20, 2019

Dr. Michael Honeycutt, Chair
Science Advisory Board
U.S. Environmental Protection Agency
Washington, DC

Re: Comments for the SAB August 27, 2019 Teleconference

Dear Dr. Honeycutt and members of the Science Advisory Board:

The Health Effects Institute (HEI) has a longstanding commitment to producing science of the highest integrity, quality, and transparency, built on a foundation of:

- Rigorous research and statistical design – subject to continuous oversight, data quality assurance audits, and more;
- Extensive efforts to test all findings against a wide range of different statistical techniques and assumptions;
- Intensive independent peer review, with *all* results published; and
- An active *Data Access Policy* for over 20 years to ensure access to underlying data for all HEI-funded studies, including:
    - Making all of the publicly available data used in the studies available for access either through the investigator or in some cases through HEI
    - Making the metadata and statistical code used in the analyses publicly available; and
    - In cases where the data cannot be released owing to existing binding privacy rules, providing detailed descriptions and/or advice on how to access private medical information which may be access through government or private resources
- HEI has also engaged on several occasions – at the requests of sponsors, Congress, and others - in gaining access to all of the data underlying a study and conducting a detailed, independent reanalysis of the study, while protecting the privacy of the individuals involved.

We are, unfortunately, unable to participate directly in your August 27 teleconference; in lieu of that, and based on our experience in producing, reviewing, reanalyzing, and interpreting science, we submit the following specific comments for your consideration:

- ***What are the protections for personal privacy?***
- ***Can the Data Still be Accessed?***
- ***Will "Depersonalized" data sets allow for full replication and reanalysis?***

1. What are the protections for personal privacy?

HEI has extensive experience with the careful and protected use of private information (residential, medical, etc.), which is critical to conducting high quality and reproducible air quality and health research. Data collected from individual subjects in a study which normally includes detailed personal, health status, and behavioral information, is important 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.

There are of course longstanding federal rules for protecting the privacy of individual information of the subjects of studies (e.g. the "Common Rule" mandating Institutional Review Board scrutiny of any use of personal data; non-disclosure of personal information through HIPAA, and others) and it is important to adhere to these even as the valuable information contained in such records is applied in scientific research.

In addition, in the case of many private and publicly-funded studies this data is normally collected through individual questionnaires and/or medical examinations, 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.

2. Can the Data Still be Accessed?

Fortunately, there are large and growing government-owned or funded data sets that can be used in these studies, and there *are* means available through a number of government agencies to make much such data available in detail to researchers, conditional on their agreeing to a data use agreement that enables access to the data – but prohibits public disclosure of individual data. For example:

- HEI-funded investigators and a number of privately-funded investigators have accessed the Medicare data sets though application to the Center for Medicare Services (CMS) and conducted and published research findings.

- Alternatively, many agencies make the data available through Federal Research Data Centers and both government- and privately-funded investigators have gained access to these data in recent years and published their results (e.g. from the National Health Interview Survey and the NCI/NIOSH Diesel Exposed Miners Study (DEMS)).

There is also the possibility – for private data sets which are not available through government portals – for new investigators to gain access to the data in collaboration with the original investigators. Such collaborations can at times be challenging to arrange, and do not necessarily allow for fully independent new analyses of the data. One alternative is for an independent third party to oversee the new analyses, similar to the process that 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.

While each of these options – and others – have strengths and weaknesses and can result in the incurring of costs, they can and have been used for an increasing number of air pollution and health studies. HEI does believe that there are improvements that could be made to the federal access options, e.g. easier access provisions for Federal Research Data Centers, and would urge EPA to work with its federal agency counterparts to accomplish that.

3. Will "Depersonalized" data sets allow for full replication and reanalysis?

Some have argued that it should be possible to create a "depersonalized" data set by stripping all personal identifiers such as address, date of birth, etc. and making such a data set widely available. However, it is not possible to conduct a high-quality air pollution and health study without knowing the locations of those being studied, i.e. where they live, and what are the sources and levels of their air pollution exposure. And unfortunately, once that information is available at smaller spatial scale, it is possible to disclose extensive medical information for individual study subjects.

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 several aspects of large population air pollution studies that have moved them towards using data at smaller and smaller spatial scales:

- First, in response to valid criticisms that the earlier air pollution 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 and smaller levels). 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 many of these studies include subjects from a wide range of community sizes, including smaller communities where identification could be possible.

- Second, as these types of 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] To address both of these issues, one of the first steps that investigators have taken has been to use data at smaller scales 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 – significantly increase the difficulty and reduce the utility 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 or perform alternative analyses. The other mechanisms discussed above – e.g. data use agreements, research data centers – fortunately would allow access to the more detailed data necessary to conduct such full new analyses while protecting the confidentiality of study subjects.

In closing, we view transparency in research is important to science, and to the credibility of studies used to support critical environmental health decisions. To that end, we feel that mechanisms can be set up to obtain and analyze personal data, but that such endeavors need appropriate institutional mechanisms and oversight.

Thank you for the opportunity to submit these comments. We would be pleased to provide any further information that the SAB might find useful.

Sincerely,

/s/

Daniel Greenbaum, President
Health Effects Institute

---

[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.