

## **Comments of the Health Effects Institute Supplemental Notice of Proposed Rule Making**

### **“Strengthening Transparency in Regulatory Science” (Proposed Rule Docket Number: EPA–HQ–OA–2018–0259)**

**May 18, 2020**

The Health Effects Institute (HEI) is pleased to submit these comments on Proposed Supplemental Rule “Strengthening Transparency in Regulatory Science.” (Docket No: EPA–HQ–OA–2018–0259). HEI appreciates EPA’s needs for ensuring the quality and reliability of the science that provide the basis for regulatory decision making. While in our comments HEI raises questions about the supplemental proposal, we would welcome the opportunity to assist EPA in making appropriate improvements to strengthen its selection and application of the best science.

HEI offers these comments on the basis of our longstanding commitment to producing science of the highest integrity, quality, and transparency, and our support for responsible efforts to enhance transparency in science. Specifically, HEI’s commitment to these principles is built on a comprehensive 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 – positive and negative -- published, and
- Periodic systematic reviews of the widest range of evidence to draw broader scientific conclusions.

HEI has, as well, implemented an active Data Access Policy for over 20 years to ensure access to underlying data for all HEI-funded studies. HEI investigators are expected to describe at the outset their plans for making the data and methods from their studies available to others at the conclusion of their research. Given the not-insubstantial cost of organizing and maintaining the data and methods, HEI has also provided financial support for such efforts. This has included freestanding databases constructed by the investigators, and efforts funded directly by HEI to make such data available, either in concert with the original investigators or independently, and posted at readily accessible data repositories.

Our extensive experience in producing, reviewing, reanalyzing, and interpreting transparent

science places high value on making underlying data available, but also emphasizes that availability of data is only one of many different aspects of each study that contribute to that study's quality and strength, and verification, should that be deemed necessary. Based on this comprehensive experience of how to judge study quality, HEI respectfully submits the following specific comments on the Supplemental Rule:

***HEI's experience makes clear that EPA should use the broadest possible range of science for making decisions on risk, causality, and other important policy. That science should be subject to critical and independent review and synthesis; it would, however, be counter to commonly accepted best scientific practice for EPA to limit the science it considers based on just one characteristic of a study.***

While access to underlying data for a particular study can be valuable, the systematic assessment of all of the available scientific evidence is the most important component of environmental policy making. Evidence integration (NRC 2014) is used by a wide variety of organizations when evaluating the information for the potential harm caused by environmental agents. Given the diverse nature of disciplines and methods that can shed light on health effects, most data synthesis procedures evaluate the findings from myriad studies using different approaches, even if they might place different weights on them. Ideally, the degree of confidence in the findings of an effect increases along a spectrum of results that progressively converge with one other, beginning with the technical and growing to a more conceptual level.

Recognizing that each discipline, method and study has its own limitations, a key step in verification of any particular finding is through "triangulation" (Cite: Lawlor, 2016, Pearce, 2019), that is, addressing the same question using multiple lines of evidence from disparate approaches to compensate for the inherent limitations in any one approach -- the more different such approaches, the better. A different population, also possibly with a different design or strategy (e.g., examination of the impact of air pollution reduction vs. observations based on existing patterns alone) is more persuasive than solely testing if the results of the original study were valid. Convergence of findings from experimental exposure of human subjects and epidemiologic findings can provide confirmation. Mechanistic insights, based on animal, cellular and in vitro studies, can provide the biological underpinnings and plausibility of clinical or epidemiological observations.

The EPA approach to evidence synthesis - before the current proposal - has usefully relied on a form of evidence integration provided by examining all evidence from a variety of sources to reach final determination of exposure and health effects (US EPA 2015). Other federal agencies, such as the Office of Health Assessment and Translation within NIEHS (OHAT, 2015), have also applied similarly comprehensive and systematic approaches.

There are, of course, opportunities to improve the systematic review processes at EPA and

other agencies, and there is active discussion about the best way to conduct systematic reviews.<sup>1</sup> However the use of evidence from the widest range of study types and designs is by far the most effective and widely accepted way to draw scientific conclusions.

While reproduction of the results of single studies is sometimes helpful in such systematic reviews, limiting evidence synthesis and assessment to only studies that have accessible data would seriously circumscribe the Agency's decision making and the robustness of such decisions. We propose below a much more effective means of selecting studies as "pivotal science" for use in policy-relevant assessments of science

***EPA has proposed two options for consideration of data and methods availability in the Supplemental Rule, both of which would restrict EPA's ability to identify and apply the best science as "pivotal science" in its regulations and assessments. While HEI agrees that such data availability is a valuable characteristic of a study, application of it as a sole or preferred criterion will significantly limit the quality of EPA's scientific review and assessment.***

In the proposed options for section 30.5 of the Supplemental Rule EPA suggests two options, that in selecting pivotal regulatory or other pivotal science the Agency will either "***only use***" or "***other things equal, give greater consideration to***" "pivotal regulatory science and/or pivotal science that includes studies with restricted data and models (*i.e.*, those that include confidential business information (CBI), proprietary data, or Personally Identifiable Information (PII) that cannot be sufficiently de-identified to protect the data subjects) if there is tiered access to these data and models in a manner sufficient for independent validation, and studies that do not include restricted data and models if the data and models are publicly available in a manner sufficient for independent validation." (emphasis added)

Based on HEI's extensive experience in producing, reviewing, and disseminating high quality science, these two limited choices could eliminate a large number of otherwise well-designed and conducted studies, and would substantially restrict EPA's ability to apply the best science to its decision making. In our view, the main criteria for selecting such studies should include:

1. Is the study representative of a broader body of evidence which, as indicated in our first comment above, has conducted a range of analyses of the same question, applying a range of analytic approaches, in a range of data sets, yet finding consistent results?
2. Amidst that evidence, does the selected study have particular strengths or attributes which suggest that it would be especially useful to be the basis of further analysis and assessment? For example, for a population study, what are the quality of individual

---

<sup>1</sup> Cf. The Problem with Mechanistic Risk of Bias Assessments in Evidence Synthesis of Observational Studies and a Practical Alternative: Assessing the Impact of Specific Sources of Potential Bias (Savitz 2018)

and covariate data available – including exposure to the pollutant, the representativeness of the population as compared to the broader US population or sensitive subpopulations, the appropriateness, quality, and innovation of the statistical methods, the extent that sensitivities have been fully tested, and other design characteristics?

3. Have the methods and data underlying the study already been made available for and subjected to efforts to independently reanalyze, reproduce, and confirm the findings of the original study and have extended or alternative analyses validated the original findings? The current supplemental rule language does not allow for consideration of any such prior confirmation and validation. (We would note, for example, that the comprehensive HEI Reanalysis of the Harvard Six Cities and American Cancer Society Studies (Health Effects Institute, 2000) would qualify as such an independent validation of the studies under any reasonable definition of such an independent effort.)
4. As one of these several criteria, are the data and statistical methods available to investigators to conduct further analyses and reanalyses? Importantly, this should recognize that in some cases, given the age of the study and the substantial changes that have occurred over time in data storage, data availability requirements, and recordkeeping rules, it may not be possible to provide access to all underlying data and methods for a pivotal study for which there are no viable alternatives.

***Although it is important to review individual studies in detail -- and examine their strengths, limitations, and ability to be reproduced -- conducting high quality efforts to reproduce results is far more complex than simply having access to the underlying data and methods. And at the same time, the proposed requirement to make all such data and methods fully available would impose substantial costs on the investigators and their institutions for which there are not readily available funds; such costs and effort are not estimated or acknowledged in the proposal.***

In addition to broader reviews of all evidence, under existing rules and by precedent EPA already has the ability and duty to assess the quality and robustness of results of each major study even in cases where the data are not available, both by careful review of all of the methods and supplemental information presented, and by expert review by EPA scientific staff and scientific advisors. These steps can identify both the strengths and weaknesses of any such study in a manner that allows the proper weighing of that study in consideration of the weight of evidence for or against a specific health effect.

In contrast, a requirement that any pivotal study shall ensure that “data and models are available in a manner sufficient for independent validation” implies a complex and potentially costly and time-consuming set of steps that include making available two detailed and complex sets of information:

- The workflow that was implemented to create the *analytical data set* used for the analysis (e.g. harmonized in time and space) from the *original and often heterogenous data sets* (individual electronic health records, gridded exposure data, confounders etc.), and
- The *data analysis* (statistical software) that was used to analyze the *analytical data set*.

Creation of both of these in a manner that enables even highly-qualified investigators to reproduce the same results is difficult and time consuming. We also know from HEI's hands-on experience that this is likely to be costly to the researchers and their institutions – both in the creation of the data and methods hubs, and in the continuous requests for further explication from those seeking the data. We would note that in contrast to the proposal's statement in response to Executive Order 13771: Reducing Regulations and Controlling Regulatory Costs: "This action is not expected to be an Executive Order 13771 regulatory action because it relates to "agency organization, management or personnel," the proposal would clearly impose substantial additional costs on a wide range of research institutions.

***While the Supplemental Rule sets the stage for extensive detailed and rigorous reanalysis, and we agree that such reanalysis may be appropriate in some cases, undue emphasis on such reanalysis is costly if done correctly and applied in all cases, and reduces resources available for new, better-designed studies that can advance scientific knowledge.***

In a limited number of cases there may not be comparable studies available in other datasets, and it could be useful to gain access to the original study data and statistical approaches to allow for independent reanalysis that asks: Can the original results be reproduced? And are they robust to a wide range of alternative assumptions, models and potential confounders? If such detailed, independent reanalysis has already been undertaken, it can significantly reduce the need for further independent validation of a specific study.

This is of course the approach that HEI applied – at the request of EPA, industry, and Congress – in its independent, rigorous reanalysis of the Harvard Six Cities and American Cancer Society Studies. The full reanalysis – which cost \$900,000 in 1998-2000 – involved data audits and quality control, replication of the original results, and extensive testing of those results against a wide variety of alternate data, assumptions, and models (Health Effects Institute 2000).

While this approach can – and did – provide comprehensive, independent assurance of the integrity and validity of the original results, it is also a highly cost-intensive undertaking and should be considered only in those exceptional cases where there is not an ability to otherwise evaluate the results of a study in the context of the wider evidence as we note above. Otherwise, we could see a substantial increase in scarce research resources and time being applied to reanalysis, rather than being invested in advancing science with new approaches and datasets; thereby further limiting the overall scientific evidence that can inform policy making.

***Although the proposed Supplemental Rule acknowledges that personal information may not be able to be protected in all cases, and therefore "tiered access" to these data may be suitable, it is critical to recognize that while "depersonalized" data sets can be created, in many instances they will not allow for full replication and reanalysis – and***

*could lead to incomplete and misleading results.*

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, medical history, 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, the sources and levels of their air pollution exposure, as well as their specific covariates, and specific health outcomes. And unfortunately, should such information be made available at smaller spatial scale, it is possible to disclose extensive personal and medical information for individual study subjects, raising privacy and ethical concerns.

Since the goal should be to find ways to share data that 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 utilizing data at smaller spatial scales:

- First, in response to valid criticisms that earlier air pollution studies relied only on central air quality monitoring data to estimate exposure, investigators have increasingly sought to improve exposure estimates by employing land use regression 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 may not allow for identification of individuals, the national analyses in some of these studies include subjects from a 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: the so-called “ecological confounding”<sup>2</sup> and “spatial autocorrelation.”<sup>3</sup> 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 address 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. Analyses using more limited data sets may well therefore result in

---

<sup>2</sup> Ecological confounding arises when some community-level variables (such as socio-economic status), which are themselves independent risk factors for mortality, are also associated with air pollution levels.

<sup>3</sup> Spatial autocorrelation is the tendency for variables to have similar values for people or areas that are geographically close, which violates a common assumption in statistical analyses that measured outcomes are independent of each other. Appropriately addressing spatial autocorrelation and accounting, for example, for other mortality causes or exposures in a given locale leads to more robust and precise results.

incomplete and misleading results.

***Finally, should EPA - despite the challenges described above – proceed with this Supplemental Rule even in a more limited form, EPA should limit the application of these rules to analyses and models developed and applied after the effective date of the rule; reaching back to every study done in the past will significantly curtail the evidence available to inform EPA decisions.***

The Preamble for the Supplement Rule states: “Proposed 40 CFR 30.5 would maintain the temporal approach to data and models taken in the regulatory text of 40 CFR 30.5 of the 2018 proposed rulemaking, and thus would apply to data and models evaluated at the time a significant regulatory action or influential scientific information is developed, regardless of when the data and models were generated. EPA is requesting comment on whether this should apply only to data and models that are generated (i.e., when the development of the data set or model has been completed or updated) after the effective date of this rulemaking.” (Emphasis added)

Although the proposed Supplemental Rule does suggest that the age of a study and the difficulty of gaining access to the underlying data could be a basis for the Administrator finding it impractical to make such data and methods available, and therefore exempting the study from this rule, such after-the-fact and potentially arbitrary determination would obstruct the continued use of a larger body of evidence – especially occupational studies that are critical to consider the effects of the toxicity of a number of specific chemicals.

EPA should recognize as well that the expectations for data access for studies have evolved over time and in many cases, older studies that did not and are unable to comply after a prolonged interval still have substantial scientific merit and value for informing regulations. Some such studies, e.g. the NIOSH/NCI Diesel Exposed Miners Study (DEMS) have made their data available through federal Research Data Centers. However, the data for other critical studies, e.g. the North American Rubber Workers Study (Delzell, 2006), which has played and continues to play a key role in assessment of risk from 1,3-butadiene and styrene, was designed and implemented at a much earlier time and would not be readily available. Yet the Supplemental Rule as proposed would require a significant delay in the use of a study that has been vetted and peer-reviewed by a wide range of scientists (including – in both these cases - the most recent update subjected to HEI project oversight and intensive independent peer review).

To ensure that EPA always has the highest quality science available for its timely and careful assessment of the effects of pollution, HEI would strongly recommend that this rule, if it were to go forward without addressing the serious concerns we note above, be applied solely to studies and analyses conducted after the effective date of the rule.

\*\*\*\*\*

In closing we appreciate the opportunity to present these comments. We firmly believe that there are ample other opportunities to enhance transparency and reproducibility in science to inform decisions, and HEI would welcome the opportunity to assist EPA in making these improvements. Should the Agency have any further questions, please feel free to contact Dan

Greenbaum, President, Health Effects Institute, [dgreenbaum@healtheffects.org](mailto:dgreenbaum@healtheffects.org), (617) 488-2331.

## **REFERENCES**

Delzell E, Sathiakumar N, Graff J, Macaluso M, Maldonado G, Matthews R. August 2006. An Updated Study of Mortality Among North American Synthetic Rubber Industry Workers. Research Report 132. Health Effects Institute, Boston MA.

(<https://www.healtheffects.org/publication/updated-study-mortality-among-north-american-synthetic-rubber-industry-workers>)

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.

(<https://www.healtheffects.org/publication/reanalysis-harvard-six-cities-study-and-american-cancer-society-study-particulate-air>)

Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol*. 2016;45(6):1866-1886. doi:10.1093/ije/dyw314

National Research Council. 2014. *Review of EPA's Integrated Risk Information System (IRIS) Process*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/18764>.

(<https://www.nap.edu/catalog/18764/review-of-epas-integrated-risk-information-system-iris-process>)

Office of Health Assessment and Translation (OHAT) Division of the National Toxicology Program National Institute of Environmental Health Sciences, Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration January 9, 2015

([https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015\\_508.pdf](https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf))

Pearce N, Vandenbroucke JP, Lawlor DA. Causal Inference in Environmental Epidemiology: Old and New Approaches. *Epidemiology*. 2019;30(3):311-316. doi:10.1097/EDE.0000000000000987

Savitz DA, Wellenius GA, Trikalinos TA, The Problem With Mechanistic Risk of Bias Assessments in Evidence Synthesis of Observational Studies and a Practical Alternative: Assessing the Impact of Specific Sources of Potential Bias, *American Journal of Epidemiology*, Volume 188, Issue 9, September 2019, Pages 1581–1585, <https://doi.org/10.1093/aje/kwz131>

U.S. EPA. Preamble To The Integrated Science Assessments (ISA). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-15/067, 2015.