April 29, 2020

Environmental Protection Agency
1200 Pennsylvania Avenue, NW
Washington, DC 20460

Dear Ms. Maria Doa,

On behalf of the Center for Data Innovation (datainnovation.org), we are pleased to submit comments in response to the Environmental Protection Agency's (EPA) request for information on the supplemental notice of proposed rulemaking (SNPRM) in "Strengthening Transparency in Regulatory Science" which seeks to ensure that the science supporting EPA's decisions is transparent and available for independent validation.[1]

The Center for Data Innovation is the leading think tank studying the intersection of data, technology, and public policy. With staff in Washington, D.C. and Brussels, the Center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as important data-related technology trends. The Center is a non-profit, non-partisan research institute affiliated with the Information Technology and Innovation Foundation.

The SNPRM significantly changes the provisions initially proposed in the Strengthening Transparency in Regulatory Science Proposed Rulemaking ("2018 proposed ruling") in three key ways. First, EPA is proposing that when making decisions, it will employ one of two approaches: to only use scientific studies where the underlying data and models are publicly available, or to give greater consideration to studies through a tiered system based on how sufficiently available the underlying data and models are for independent validation. Second, the SNPRM increases the scope of the 2018 proposed ruling to apply to all scientific data and models, not only dose-response data and models (which are those that describe the effect of increased levels of exposure of an agent on an organism or the environment). Third, the SNPRM increases the scope of the 2018 proposed ruling to apply to all decisions related to influential scientific information, i.e. those that have impact on public policies or private sector decisions, not only significant regulatory decisions.

By introducing public availability of data as a measure for the integrity of scientific evidence, EPA is reserving the right to itself to give less weight—to the point of no weight whatsoever— to important

[1] "Strengthening Transparency in Regulatory Science," Environmental Protection Agency, April 13, 2018, https://www.regulations.gov/document?D=EPA-HQ-OA-2018-0259-9322

scientific studies which do not make their underlying data publicly available for various reasons. While open access to scientific research is a laudable goal, regulators should not ignore legitimate research simply because the data is not public. Severely limiting the types of research EPA considers when developing critical public policy only serves to diminish its ability to protect public health and safety.

DIGITAL INFRASTRUCTURE OVERCOMES BARRIERS TO TRANSPARENCY

EPA should consider the entire body of cumulative evidence on any given subject and implement policies that aid in overcoming technological barriers to reproducibility rather than excluding studies. Specifically, we recommend EPA invest in the research and development of open-source, usable tools, and infrastructure that produce statistically sound, reproducible, and verifiable results. Scientific research today is largely computationally heavy and follows a workflow, such as the intricate numerical modeling of climate and weather. Scientists need standards-based workflow systems that better document, track, and detail their research and enable them to map the complex process from data to results.[2] This will make it easier for other researchers to independently verify any claims. Instead of enforcing strict transparency rules on the underlying data—which cannot always be transparent either to protect the privacy of identifiable data or to protect the intellectual property of proprietary data—EPA should encourage transparency through initiatives aimed at methodology.

Further, we support the recommendation from the National Academic of Sciences, Engineering, and Medicine (NASEM) that the National Science Foundation (NSF) consider creating code and data repositories for the long-term archiving of digital artifacts and believe EPA should endorse this.[3] Not being able to access legacy data in a usable format presents an obstacle for reproducibility. What is needed are data archives that provide long term stewardship such as the U.S Department of Energy's (DOE) Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE). This data repository stores, expands access to, and improves usability of data from DOE's research in terrestrial and subsurface environments.[4] Researchers in all fields of environmental studies could benefit from such resources and EPA could benefit from the increased openness of their data.

---

[2] Yolanda Gil et al., "Examining the Challenges of Scientific Workflows," *Computer* (December 2007), 24-30. https://doi.org/10.1109/MC.2007.421

[3] National Academies of Sciences, Engineering, and Medicine (NASEM), *Reproducibility and Replicability in Science*, (Washington, DC: The National Academies Press 2019). https://doi.org/10.17226/25303

[4] "ESS-DIVE, Environmental Systems Science Data Infrastructure for a Virtual Ecosystem", accessed April 11, 2020, https://ess-dive.lbl.gov/

REPRODUCIBILITY ALONE IS A POOR INDICATOR OF QUALITY

In the new rules, EPA presents public availability of a study's underlying data as a proxy for how reproducible a study is. EPA predominantly focuses on reproducibility as a gauge for the validity and reliability of scientific studies and therefore contends reproducibility, by way of data transparency, is the most important factor in judging scientific knowledge. But this contention is misguided.

There is no doubt that reproducible studies instill credibility in scientific knowledge, but reproducibility is an inadequate measure to solely judge the rigor of scientific research. Successfully reproducing a result does not prove that the original scientific result is correct, nor does failing to reproduce it definitively negate the original claims. In fact, the less variation there is in the conditions of a study, the fewer contexts in which the results are true.

For instance, a 2009 study on the effect of environmental variations on reproducibility in animal experiments found that reducing variation led to results that were only true in the specified conditions, with little external validity.[5] When the extent to which the results of a study do not apply in other contexts or populations that differ from the original one, the results cannot be determined to be robust enough, no matter how reproducible they are. EPA should not be focusing solely on reproducibility to ensure data is valid, but instead should concentrate on generalizability as well as the effect and size of data.

We recommend EPA revokes its proposal to prioritize scientific studies on the nonscientific basis of public availability of underlying data. This will allow EPA to assess the entire body of research on a subject and gain confidence in the state of scientific knowledge by more accurate benchmarks of quality; namely generalizability, not reproducibility.

EDUCATION ENCOURAGES OPENNESS IN RESEARCH CULTURE

Regarding EPA's request for comments on how best to incentivize researchers to increase access to data, we suggest that lowering barriers associated with making data available is a better mechanism for change. According to the 2019 NASEM report, the lack of a supportive framework for adequate recordkeeping disincentivizes researchers to create the conditions for reproducibility.[6] EPA can address this not only by fostering the development of new software tools, but by providing adequate

[5] S Helene Richter, Joseph P Garner and Hanno Würbel, "Environmental Standardization: Cure or Cause of Poor Reproducibility in Animal Experiments?", *Nature Methods*, no. 6 (2009): 257-261, https://www.nature.com/articles/nmeth.1312

[6] National Academies of Sciences, Engineering, and Medicine (NASEM), *Reproducibility and Replicability in Science*, (Washington, DC: The National Academies Press 2019). https://doi.org/10.17226/25303

training on how to use such tools and education in the importance of transparency in scientific studies.

This can be achieved by partnering with educational institutions and leveraging the working relationships EPA's Office of Public Engagement and Environmental Education (OPEEE) already has, such as with the U.S. Department of Education, to create programs that promote a culture of openness in the scientific community.[7] An example is the "Reproducible Research" course at John Hopkins University which focuses on the concepts and tools behind reporting data analyses in a reproducible manner.[8]

DATA SHOULD NOT BE EXCLUDED WHEN MAKING POLICY DECISIONS

Given the proposal in the SNPRM to expand the applicability of the ruling from dose-specific data to all influential scientific information, the implications of ill-informed policy are even more wide reaching. The rule now extends its scope to include bioaccumulation data, water-solubility studies, environmental fate models, engineering models, data on environmental releases, exposure estimates, quantitative structure activity relationship data, and environmental studies.

These data inform policy decisions around public health and environmental policy such as the 2011 Mercury and Air Toxics Standards (MAT). This regulation, based on bioaccumulation data of mercury, saves up to 11,000 premature deaths in the United States according to the EPA's own projections.[9] The proposed rule, as clarified by the SNPRM, would undermine EPA's fundamental mission to develop and enforce regulations that protect human health and the environment.[10]

The proposed rule would impede EPA from using relevant scientific data in the hundreds of decisions it makes every year. We strongly recommend that EPA does not use its authority independently or in conjunction with environmental statutory provisions to take the actions in the proposed ruling. We suggest EPA withdraw the proposed rule and instead adopt policies that overcome the technological and cultural barriers to data sharing and reproducibility.

---

[7] "About the Office of Public Engagement and Environmental Education (OPEEE)," last modified March 24, 2020, https://www.epa.gov/aboutepa/about-office-public-engagement-and-environmental-education-opeee

[8] Johns Hopkins Bloomberg School of Public Health OpenCourseWare website, accessed April 10, 2020, https://ocw.jhsph.edu/index.cfm/go/viewCourse/course/repdata/coursePage/index/

[9] "Healthier Americans," last modified December 7, 2016, https://www.epa.gov/mats/healthier-americans

[10] "About EPA: Our Mission and What We Do," last modified on February 7, 2018, https://www.epa.gov/aboutepa/our-mission-and-what-we-do.

Sincerely,

Daniel Castro
Director
Center for Data Innovation
dcastro@datainnovation.org

Hodan Omaar
Policy Analyst
Center for Data Innovation
homaar@datainnovation.org