Comment on the Supplemental to Strengthening Transparency in Regulatory Science

May 7, 2020

Prepared with the input and guidance of the ASA Privacy and Confidentiality Committee and the ASA Scientific and Public Affairs Advisory Committee

Founded in 1839, the American Statistical Association is the oldest scientific professional association in the United States. With 18,000 members in academia, industry, and government and statistics being the science of learning from data, ASA's membership and expertise is especially diverse. ASA member expertise covers a myriad of topics including, for these comments, privacy and confidentiality, reproducibility, data analysis, optimization of the scientific process, experiment and survey design, decision analysis and support, minimization of and accounting for bias, and uncertainty quantification.

We are grateful for this opportunity to comment on the EPA's supplemental to its 2018 proposed rule, Strengthening Transparency in Regulatory Science. In our 2018 comments, we urged the rule not be adopted for the following six reasons:

1. Proposed rule hampers the use of evidence.
2. Discriminating what science to use for drafting regulation risks introduces potential bias in the rulemaking process.
3. Providing access while protecting confidentiality is challenging and complicated.
4. New studies may be discouraged for privacy and/or cost considerations.
5. Costs and benefits are not fully considered.
6. Rescinding regulations may have detrimental health and environment effects.

We appreciate the EPA's efforts in the supplemental and acknowledge advances that address comments like those above. Indeed, we believe EPA's efforts to provide access to EPA-owned data and to facilitate access to outside research and data that informs EPA work are laudable and should be continued within current EPA rules. Regrettably, however, despite improvements in the proposed rule, we maintain our position that the rule should not be adopted for the following reasons:

1. Proposal to provide greater consideration to some studies has numerous weaknesses: (i) processes unspecified; (ii) standards and guidance missing; (iii) connection to standard scientific analysis method lacking; (iv) bias-minimization procedures omitted; and (v) transparency provisions left out.
2. EPA's world-class process for evaluating scientific evidence is encumbered and fettered to weaker scientific criterion and a scientifically unjustified assumption.

3. EPA's data-sharing infrastructure not in place to support the rule's objectives.

The proposed rule's defects are so fundamental that the proposed rule cannot be fixed without a complete reanalysis and a new proposal. For these reasons, we strongly caution against implementation of this proposed rule. We elaborate on these issues in the following.

1. **Ambiguity of terms and lack of established process:** The supplement introduces a new process that lacks description and that uses language requiring judgement calls for which no protocols are provided. Specifically, the supplement calls on EPA staff to determine, for example, what studies are "sufficient for independent validation" and what studies should be given "greater consideration."

   Starting with the first phrase, "sufficient for independent validation", the EPA's supplement does introduce a definition for the term, independent evaluation: "the reanalysis of study data by subject matter experts who have not contributed to the development of the original study to demonstrate that the same analytic results are capable of being substantially reproduced." However, the concern is that there is insufficient guidance or standards for what the terms, "sufficient" and "substantially", mean or who will make the determination (e.g., scientific staff with oversight of an EPA scientific advisory panel). To its credit, EPA is using the 2002 OMB definition for "Capable of being substantially reproduced". The definition however uses the words, "similar" and "acceptable," which are subject to the same concerns as "sufficient" and "substantially". It is also worth noting the extensive work on the issues of reproducibility and replication done since 2002, some of which is cited in the proposed rule and supplemental. Standards for these terms and determination should be included before this proposed rule is finalized and should include sufficient guidance to, for example, take into account how unlikely it is for two studies to exactly replicate. In 2015 Ralph Cicerone, then president of the National Academy of Sciences (NAS) addressed this point in a speech, noting that a NAS panel had grappled with the very question: "Because variability across studies is to be expected, how can we assess the acceptable degree of variability and when should we be concerned about reproducibility?" There should also be guidance to require explanation of whether research has or has not met the sufficient-for-independent-validation criterion.

   For the term, "greater consideration," the supplement does require "a short description of why greater consideration was given" when a study is given greater consideration. However, there is no (i) guidance for what "greater consideration" means when weighting the results of one study with that of another; and (ii) direction that EPA must explain the weighting given to various relevant studies. Such transparency here and in the previous paragraph is critical to rigorous scientific consideration of the relevant evidence.

   How to weigh the consideration of studies is at the heart of meta-analysis, a study that uses scientific and statistical methods to combine results from multiple studies. As we noted in our 2018 comments, published meta-analyses have grown immensely over the years (to more than 250,000) to become an important and standard research method in epidemiology, medicine, and public health. For this active and important research field, the question of what studies are of high quality is an ongoing challenge because of the recognition that no study is universally "good".

The fact that the supplement does not reference the well-known and valued scientific practice, which EPA staff regularly uses, reinforces the point that EPA has much work to do in laying the foundation for accomplishing the stated goal of the proposed rule.

These two issues—ambiguity of terms and lack of process and transparency—risk the introduction of bias, if not improper influence, into the EPA process. Specifically, as we discussed in the second reason from our 2018 comments, the two issues risk "arbitrary and capricious decisions as to what evidence to include or exclude" and "would also often result in regulatory analyses that suffer from biased selection and inadequate statistical power," among other issues.

Before the EPA implements the rule, it should design and explain its process for considering different studies. It should also acknowledge the risk of bias in the process and explain how it will work to minimize such risk. A highly regarded site is that of the Cochrane Collaboration's revised risk of bias framework available at the following url: https://www.riskofbias.info/.

We acknowledge and thank the EPA for providing definitions in the supplement for the following terms in response to comments submitted in 2018: "Capable of being substantially reproduced", "Data," "Independent validation," "Influential scientific information, " "Model," "Pivotal science," "Publicly available" and "Reanalyze."

2. **"Independent validation" criterion is counterproductive to EPA consideration of the best science:** Independent validation as the EPA defines the term in the supplement is a rather narrow method of evaluating research and assessing a scientific conclusion that could ultimately hamstring the work of EPA scientists to determine the best science to inform EPA work.

   The focus on independent validation puts undeserved emphasis on an individual study. Such a focus is contrary to the science community's practice of generally not being interested in a specific study but rather on its underlying hypotheses. No hypothesis can be conclusively addressed by an individual study. Instead, a hypothesis' validity should be assessed on a body of evidence.

   The proposed rule also seems to assume that a study that can be independently validated is of higher quality. However, there is no evidence to support such an assertion. Leek and Peng discussed in further depth the lack of a connection between research that has been reproduced and quality in a 2015 article in *Proceedings of the National Academy of Sciences*. The EPA also acknowledged this point in the afore-mentioned document, Plan to Increase Access to Results of EPA -Funded Scientific Research: "Whether research data are fully available to the public or available to researchers through other means does not affect the validity of the scientific conclusions from peer-reviewed research publications."

   A broader approach to ensuring valid findings, especially in complex areas like regulatory environmental science, is needed. Rather than focusing on single studies, evidence experts look at bodies of evidence and seek triangulation of data from a diverse set of research approaches to evaluate consistency of results and enhance confidence in findings above that of any individual study.

   As a regulatory agency, EPA work is constantly under scrutiny, putting its scientific justifications under a microscope with legal challenges an ever-present possibility. As a result, the EPA, with

the expertise of its scientific and other professional staff and with the consultation of its scientific advisory panels, has an effective and advanced system of scientific checks and considerations that have evolved and improved over the decades. As an example of EPA's rigorous scientific processes, consider the rigor, detail, and transparency of the EPA program for managing the quality of its environmental data. Consider also the EPA Integrated Science Assessments which rate evidence for certain hypotheses on a causal spectrum and carefully consider a body of evidence without undue emphasis on any one study. This rule reverses that progress by having it revert to a more narrow check and also freezes EPA's scientific validation process around this narrow check.

The rule should be withdrawn so that EPA scientists can use the most recent and rigorous methods for determining the best science to use in supporting the EPA's mission to protect the environment and the health of the U.S. population.

3. **Data sharing infrastructure not in place:** The assumed or proposed models and mechanisms for data sharing in the EPA's proposed rule are still in the early stages of development and lacking proof of concept. Therefore, they could not effectively support the aim of the proposed rule. We explain our reservations for three main classes of data that could inform EPA actions:
    a. **Data owned by the EPA or other federal entity**: For this category of data, the EPA says the data will be made "available through tiered access when the data includes confidential business information (CBI), proprietary data, or personally identifiable information (PII) that cannot be sufficiently de-identified to protect the data subjects" and gives the Research Data Center (RDC) of the National Center for Health Statistics as an example. The federal statistical agencies are indeed at the forefront of implementing tiered access for their data having provided tiered access to some microdata for decades. However, the push now is to provide expanded access to data, as required in the 2019 Evidence Act and an effort which is still in its infancy. OMB Memorandum M-19-15 acknowledges this with its Implementation Update 3.5 guidance to "explore methods that provide wider access to datasets" and statement that "tiered access offers promising ways." To this end, we also note that OMB has yet to release its expected guidance for tiered access. Consider also the September 2019 workshop on tiered access hosted by the Council of Professional Associations on Federal Statistics. The workshop's recommendations for OMB include encouragement of experimentation as well as further research and development.

    The NCHS RDC is one example for providing some level of tiered access. The NCHS also, along with more than a dozen other federal entities, provides data access through Federal Statistical Research Data Centers (FSRDCs). The EPA actually provides access to its Pollution Abatement Costs and Expenditures (PACE) survey through the FSRDCs. One potential limitation of the FSRDC network (see FSRDC requirements) is that work conducted therein must be for a statistical purpose[1] and must be for research as well as

[1]The term, statistical purpose, in this context is defined as "the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups; and includes the development, implementation, or maintenance of methods, technical or administrative procedures, or information resources that support such purposes."
https://www.federalregister.gov/documents/2014/12/02/2014-28326/statistical-policy-directive-no-1-fundamental-responsibilities-of-federal-statistical-agencies-and

the following stipulation: "An agency with regulatory or enforcement roles that also conducts or supports research may be able to house their data in RDCs if there is a clear delineation between enforcement and research in the agency. Further, the mission of RDCs is to facilitate research." Other existing access methods may provide additional flexibility and are actively being used by federal agencies.

Two such access examples are the NORC Data Enclave and the Inter-university Consortium for Political and Social Research (ICPSR) Virtual Data Enclave. The former is used by the two USDA statistical agencies, for example, and the latter by the Bureau of Justice Statistics. If more restricted access is needed, programs such as the Bureau of Economic Analysis' Special Sworn Researcher Program provides access to highly confidential firm data, under tight security.

In short, while EPA is actively working to provide more access to its data, its involvement in the sharing of its sensitive data through research data centers and data enclaves is minimal, limited to a single survey to our knowledge.

Given the EPA's limited involvement with RDCs and data enclaves and no public plans for further involvement, it is premature to enact this rule. While we cannot support this rule as modified in the supplemental, we support the idea that EPA expand tiered access to its data through the RDC's, data enclaves, or by other means.

b. **Federally funded non-EPA data:** Making federally-funded research available for which EPA does not own the data is also a work in progress as the supplement's wording states: "Development of standard data repositories is still ongoing." The supplement also notes the White House Office of Science and Technology Policy recent call for comments on a "draft set of characteristics of data repositories used to locate, manage, share, and use data resulting from federally-funded research." For a snapshot of the status of data repositories for federally funded research, consider the National Institutes of Health (NIH), the federal government's likely leader in federally funded data sharing, an effort started more than a decade ago. In its frequently asked questions (FAQ) page for data sharing,[2] the NIH responds, "maybe," to question #20 on whether researchers should consider contributing their data to a data archive. The reply equivocates of course because of the many factors involved but also indicates how far the NIH is from achieving full data sharing for the research it funds. For the next question of where to find guidance on preparing data for sharing and archiving, the NIH FAQ does not provide its own guidance but, except for molecular biology information, refers readers of the guidance to an outside entity, the afore-mentioned ICPSR. EPA is a relative newcomer to encouraging data sharing and, while it will benefit from the advances and experiences of other federal entities, is still likely to be years away from providing or encouraging the data access assumed in this rule.

We also note that, unlike the NSF and NIH, the EPA does not seem to require a data management plan of its grantees, which it planned to do by 2018 in the 2016 EPA

[2] https://grants.nih.gov/faqs#/data-sharing.htm

document, Plan to Increase Access to Results of EPA-Funded Scientific Research.[3] NIH, in fact, is seeking to expand its policy on data management and sharing (84 FR 60398), with the goal of making results more transparent to the public. Such a requirement might be a first step towards the eventual goal of EPA funded research data being more widely accessible. And while data management plans and sharing strategies remain work in progress, they are a mechanism to ensure that large swaths of data become accessible to as broad a public as possible, including through some of the same restricted-access mechanisms mentioned in the previous section.

Without ready access to this category of data by a known date in the foreseeable future, we believe it is premature to enact this proposed rule. We do, however, urge EPA's continued work to make such data available while also respecting and ensuring the confidential aspects of the data.

c. **Non-federally funded and non-EPA data:** This data category likely makes up the largest category informing EPA's many actions. It is also the category that poses the most challenges for making data available. Such studies, however, generally provide important scientific contributions and should not be discounted because their data are not readily available. In our 2018 comments, we included two examples of research that might be excluded: private-sector studies and meta-analyses. While such work will not necessarily be excluded through the proposed rule with the supplemental, it is likely to be de-emphasized to the detriment of evidence-based policymaking at EPA.

Meta-analyses and systematic reviews have become important tools in evidence-based policy making. Meta-analyses present a special challenge for data sharing because meta-analyses often do not begin with raw data, but with data that have already been "processed" to some degree (for example, the data may be from the results of primary analyses in the studies contained in the meta-analysis). In such cases, the definition of raw data underlying the meta-analysis is ambiguous and the author of the meta-analysis may not have access to the underlying data in each study used in the meta-analysis. Even if the data were available for meta-analyses, data repositories for meta-analytic data do not currently exist. Systematic reviews, one particular form of meta-analysis, are regularly used by EPA to bring together the many relevant studies and develop a summary weight-of-evidence recommendation.[4,5] De-emphasizing such analyses could greatly weaken EPA's ability to develop regulations and make them harder to defend in court challenges.

We emphasize the extent to which researchers go to protect the privacy of the individuals' whose data may be part of the research. For example, researchers are developing multi-level models and privacy-preserving distributed algorithms to draw scientific conclusions based on data stored in multiple locations. Related work is also a

[3] The most recent update on this plan is from 2018: https://www.epa.gov/sites/production/files/2018-07/documents/reportonepaplantoincreaseaccesstoresearch.pdf.

[4] For systematic reviews, see, for example, https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/application-systematic-review-tsca-risk-evaluations.

[5] For meta-analysis, see, for example,https://www.epa.gov/environmental-economics/report-epa-work-group-vsl-meta-analyses-2006.

major focus in medical informatics for data from different hospitals (e.g., electronic health record data[6] and biobank data[7]) while protecting patient privacy. Making such data available in any way would be impractical if not impossible.

More generally, research that uses CBI, proprietary data, or PII is likely to provide greater insights into health effects pertinent to the EPA's work precisely because of these data. Because such data will not be "publicly available or available through tiered access," the potentially more insightful research that relies on these data is likely to be discounted by this proposed rule to the detriment of EPA's important work.

For CBI, for example, EPA uses such data to identify best available technologies (BAT) upon which many water regulations are based. Companies only provide the data upon which BAT can be determined with the proviso that the data remain CBI. Without accurate data on the BAT currently in use, such regulations would be based on weaker science. The proposed rule poses three risks for this example: (i) businesses may have less confidence their CBI will be protected and may be reluctant to share their information; (ii) EPA may discount research based on CBI weakening the scientific underpinning of a regulation; and (iii) water quality may suffer.

We therefore do not believe scientific reasoning would support the following aspect of the supplement: "other things being equal, the Agency will give greater consideration to studies where the underlying data and models are available in a manner sufficient for independent validation either because they are publicly available or because they are available through tiered access when the data includes CBI, proprietary data, or PII that cannot be sufficiently de-identified to protect the data subjects." In short, EPA has not scientifically justified that greater consideration should be given to such studies.

The reasons stated above reiterate some of the six concerns from our 2018 comments. Specifically, we believe these three concerns will hamper the use of evidence in EPA work because of the discounting of important research and introduce potential bias in the rulemaking process, which in turn may have very detrimental health and environment effects. We also maintain that costs of providing data access are not fully considered. The costs of data sharing are not negligible and the community has yet to fully account for such costs. Indeed, the NIH data-sharing FAQ addresses the concern as does the U.S. Census Bureau guidance for establishing an RDC. This rule should not be finalized without a full cost-benefit analysis as well as an explanation for how the rule implementation and execution will be funded.

To summarize, EPA should have an operations plan and an operational infrastructure for data sharing before implementing the requirements in the proposed rule as updated by the supplement. The operations plan should address financial support and logistics for placing research data into an enclave or RDC and carrying out the reanalysis and independent validation.

---

[6] R. Duan et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm, Journal of the American Medical Informatics Association, Volume 27, Issue 3, March 2020, Pages 376–385, https://doi.org/10.1093/jamia/ocz199.

[7] R. Li., Y Chen, M.D. Ritchie et al. Electronic health records and polygenic risk scores for predicting disease risk. Nat Rev Genet (2020). https://doi.org/10.1038/s41576-020-0224-1.

For the EPA to continue improving access to its data and the data it uses, it may be helpful to consult the insightful, constructive and comprehensive 2019 Bipartisan Policy Center technical paper, Meaningful Transparency at EPA: A Framework for Rationalizing Approaches to Promote Open Science and Data Sharing for Evidence-Based Policymaking.

It is admirable for EPA to work toward reproducibility of research sponsored by EPA, and a data governance system that provides tiered access to confidential data is an appropriate way to achieve this. Nevertheless, EPA would do itself and the country a great disservice to exclude or discount important research studies conducted under different laws and policies than its own, because they may not meet the standards imposed by the rule.

EPA can ensure that over time more of the data and models underlying the science that informs significant regulatory decisions are available publicly or via restricted access by working together with other agencies to develop common standards and provide resources to researchers and research sponsors to create data governance systems that support tiered data access. In the meantime, we believe EPA should postpone the implementation date of the ruling to evaluate the obstacles to meeting the rule requirements for research not sponsored by EPA and the impact of the rule on policymaking. It is incumbent on the EPA to ensure that the requirements of the proposed ruling can be met by the vast majority of research used by EPA prior to implementation of this rule.