SILENT SPRING INSTITUTE
Researching the Environment and Women's Health

320 Nevada Street, Suite 302, Newton, MA 02460 / 617.332.4288 / www.silentspring.org

May 15, 2020

Andrew Wheeler
Administrator
Environmental Protection Agency
1200 Pennsylvania Avenue, NW
Washington, DC 20460

Submitted electronically via regulations.gov

Re: Strengthening Transparency in Regulatory Science (Docket No. EPA-HQ-OA-2018-0259), Supplemental Notice of Proposed Rulemaking

Silent Spring Institute is an independent research organization committed to investigating links between the environment and women's health, with a focus on breast cancer. It was founded as a collaboration of scientists, clinicians, and families affected by breast cancer, with a mission to conduct scientific research that can inform disease prevention. As scientists and as partners with women across the country diagnosed with breast cancer, we urge U.S. EPA to use the best available science to set environmental regulations that protect health.

As we described in a prior comment (EPA-HQ-OA-2018-0259-6352), the proposed rule "Strengthening Transparency in Regulatory Science" created a number of barriers to using the best science to protect public health. Although EPA intended to address some of those barriers in this Supplemental Notice of Proposed Rulemaking, we remain concerned that this rule will (1) cause EPA to ignore or diminish the impact of high-quality and irreplaceable scientific research and (2) discourage public participation in future environmental health studies due to privacy concerns. We also comment closely on the proposed system of "tiered access," which has the potential to leave participant privacy unprotected if not implemented broadly.

Tiered access is a critical new feature that is poorly defined by the Supplemental Notice

We are relieved to see EPA acknowledge that some data cannot be made publicly available without violating participant privacy. In an effort to address this problem, the Supplemental Notice proposes a system of tiered data access, whereby data containing personally identifiable information (PII) would be accessible to authorized researchers through secure data enclaves. However, we have a number of concerns about how the tiered access system will be implemented that were not addressed in the Supplemental Notice.

1. What data are eligible for storage in the restricted access data enclave?

According to the Supplemental Notice, the restricted enclave is intended for "data and models that include CBI, proprietary data, or PII that cannot be sufficiently de-identified to protect the data

subjects." Our focus is data collected about human subjects. A primary concern about data transparency is that study participants will be re-identified—when identifiable information sufficient to contact a person (such as name or address) is matched to "de-identified" data. The Supplemental Notice defines PII by reference to the "requirements of the Common Rule, the Health Insurance Portability and Accountability Act (HIPPA) [sic], the 21st Century Cures Act, the Privacy Act, and other relevant laws and regulations, and EPA privacy policies." **This definition of PII does not capture all of the data types that can contribute to re-identification risk.**

Linkage re-identification can occur when de-identified data is matched to externally available, identifiable data using data fields that overlap between both datasets. In NIH-supported research recently published in *Environmental Health Perspectives* (1), we analyzed twelve prominent environmental health studies and identified five different data types—location, medical, genetic, occupation, and housing—that overlap with outside databases and could contribute to the risk of re-identification. We found that all 12 studies included at least two out of the five data types, and three studies included all five.

Examples of external public or commercial datasets that can be used in linkage re-identification include tax assessor data, registries of licensed professionals, hospital discharge records, or advertising lists (see Boronow et al. 2020 for a complete review, attached). Because environmental health studies often link protected health information and non-protected fields, seemingly non-sensitive data can compromise the security of the entire dataset and lead to breaches of sensitive personal information.

In an earlier experiment conducted under IRB approval, we empirically demonstrated the requirements of the HIPAA Privacy Rule, which is mentioned in the Supplemental Notice, do not protect against linkage re-identification (2). Using data from Silent Spring Institute's Northern California Household Exposure Study (3), we shared a "de-identified" version of the dataset with researchers skilled in re-identification techniques. The dataset included environmental chemical measurements in indoor and outdoor air, housing characteristics, some personal behaviors, and basic demographics. It was redacted to comply with HIPAA Safe Harbor requirements, and some non-HIPAA-protected data fields were also modified (e.g., year house built, individual room dimensions) in an effort to limit identifiability while retaining the utility of the data for analysis. The re-identification strategy relied on linking housing and demographic information to publicly-available tax assessor data. It also used information from the published peer-reviewed journal article about the study, which disclosed the housing developments where the study took place and the range of chemical levels detected in two different cities. As holders of the original data, we then scored the accuracy of their re-identifications and found that 25 percent of participants from one housing development were correctly and uniquely identified by name (2). This level of privacy risk is not ethically acceptable.

Although certain data types pose known threats to privacy, we further recognize that privacy is a moving target: as new data types are collected and new analytical methods are developed, novel privacy risks emerge that cannot always be anticipated. Mobility traces (high-resolution location information) are one example of an emerging data type in environmental health studies. These data were not commonly collected at the time when we identified the twelve studies in our earlier analysis, but they are increasingly being used to infer co-located environmental data as technological and cost barriers have decreased. These data are known to be extremely vulnerable to re-ID (4-6).

We also demonstrated how chemical exposure measurements, a distinguishing data type of environmental health studies, could pose an unanticipated privacy risk by applying a novel analytical approach. Unlike the data types discussed above, chemical exposure measurements alone are less vulnerable to data linkage because there are few databases that include chemical measurements that could be used for matching. To explore a different way that chemical exposure data might be used in re-identification, we conducted a cluster analysis using data from Silent Spring Institute's Household Exposure Study in California and in Massachusetts, and from the U.S. Department of Housing and Urban Development (HUD)/Centers for Disease Control and Prevention (CDC) Green Housing Study in Boston and Cincinnati (1). We used K-means clustering, a common algorithm for unsupervised clustering, to blindly classify participants from each study into two groups based solely on their raw chemical measurement data. The groups created by the algorithm corresponded to geographic location with 80 to 98 percent accuracy (1). Using the chemical measurement data to infer location can enhance the probability of success of a linkage re-identification attack using other fields in the dataset, because each group can be matched to data narrowed to that location, making it more likely for a re-identification attack to produce correct matches. **This analysis demonstrated that chemical data could be used to infer a characteristic of people in a study, even if that characteristic is excluded when the study data are shared.** Although our analysis was focused on inferring location, the technique could also be used infer other characteristics that co-vary with exposure, such as gender, race/ethnicity, or occupation.

Our research shows that the diverse data types collected in environmental health studies can create re-identification risk, new and risky data types can emerge over time, and even seemingly low-risk data types, such as raw chemical exposure measurements, can be used to infer latent subgroup information. Together, these findings demonstrate that it is not feasible to distinguish between data types that pose re-identification risk and those that do not. **Thus, we urge EPA to treat all data associated with human study participants and their homes as potentially containing personally identifiable information. To adequately protect privacy, all data associated with human study participants, regardless of the variables collected, should only be accessible through restricted-access enclaves.**

2. Who will be able to access restricted data?

The Supplemental Notice described the restricted access tier as "limited to authorized researchers and not possible for the general public." However, EPA does not define a process for vetting applicants or criteria for granting access. Restricted access must be limited to researchers with a legitimate scientific interest in the data, otherwise restricted access is simply 'accessible to the public by application.' The Research Data Center (RDC) run by the National Center for Health Statistics (NCHS)/CDC—identified as a possible model by the Supplemental Notice—requires researchers "submit a research proposal outlining the need for restricted-use data." However, a stated primary goal of the Supplemental Notice in making data available is to support reanalysis of study data. It is unclear on what basis applications seeking to perform reanalysis would be judged, because reanalysis is not typically considered research. EPA must clearly define how applications to access restricted-access data will be evaluated.

An established practice for evaluating applications for controlled data is to have a Data Access Committee, rather than leave decision-making in the hands of a single individual. For example, large intramural research studies conducted by the National Institute of Environmental Health Sciences (NIEHS), including the Sister Study and Agricultural Health Study, have committees to review data requests (7, 8). We advise EPA to establish a Data Access Committee that includes equal representation

by academic scientists, industry scientists, public-interest researchers, and community members who are nonscientists. Similar to requirements for IRB membership (21 C.F.R. §56.107), the committee should have diversity of membership (including race, gender, and cultural background), and members must not have any conflict of interest with the applications they are evaluating.

We also wonder how EPA intends to address issues of ownership and responsibility for data deposited in the restricted enclave. Stewardship of data involves maintenance of the infrastructure for storing and accessing the data, accountability for the data quality and underlying methods that generated it, and moral responsibility to the study participants. Researchers may be reluctant—or unable—to cede moral responsibility and accountability for the data to EPA, even if they can cede maintenance of the data. In light of their ongoing responsibilities, the original study investigators should be granted some input over who is allowed access to their data and for what purposes. One way to do this would be to give the original study investigators an automatic seat on the Data Access Committee for any application seeking to use their data. In this way study investigators will have the opportunity to review and comment on research proposals seeking to use their data.

3. How will researchers access data stored in a restricted-access enclave?

The Supplemental Notice identifies the NCHS/CDC RDC as a possible model for EPA and states that EPA is conducting a pilot study storing EPA datasets in an RDC. **We request that the details and results of that pilot study be made public.**

A key feature of the RDC process is that researchers do not have access to the restricted data except while on site at the RDC, nor are researchers allowed to bring any external technology or datasets into the RDC. All computer code for running analyses at the RDC is reviewed in advance, activity inside the RDC is monitored, and the analysis output is reviewed again before being released back to the researcher. This process ensures that researchers never have access to the restricted data except under close supervision, which precludes any possibility of a linkage re-identification attack.

This close oversight over the data comes at the cost of operating the RDC. While we do not have information on the cost to CDC of running the RDC, we do know that the RDC charges researchers accessing the data thousands of dollars in administrative fees to help support the centers' operation. Researchers are also responsible for the travel expenses associated with visiting the RDC in person. The Supplemental Notice does not detail the costs of establishing and maintaining an RDC-type system—which we anticipate to be significant—nor does it detail how EPA will fund such a system and what costs, if any, will be passed on to the users of the system (both researchers depositing data and researchers accessing data).

In the case of the RDC, the data stored belongs to NCHS (or in some cases to other operating divisions of the Department of Health and Human Services [DHHS]). However, the system employed by EPA will house extramural data produced by academic researchers in addition to data from their own scientists. We anticipate that the scope of data to be stored in the RDC—essentially all epidemiologic research on environmental exposure—to be potentially massive. Before requiring such a system, EPA must clearly explain how it will fund the staff and infrastructure required to collect, process, store, and mediate access to this immense trove of data. EPA must also define the burden to researchers of preparing data for entry into the repository, if such preparation is required. To ensure equitable use of the RDC, EPA should not charge researchers to deposit or store their data in the RDC. We are concerned that the costs

of either depositing extramural data into an RDC or accessing data stored in an RDC will disproportionately impact certain types of research interests.

An alternative model, in which researchers are given unsupervised access to the raw restricted data, depends on the integrity of researchers and a system of penalties and enforcement for misuse. We do not recommend this model.

Another alternative is to work within current practices for data sharing. Current practices for sharing individual-level participant data between researchers involves ethical oversight by Institutional Review Boards (IRBs). IRB oversight ensures that participants' rights are respected and that data are used for legitimate purposes, and it creates a record of persons having had access to the data. The National Institutes of Health and scholarly journals already encourage data sharing to maximize scientific benefit and ensure rigor and reproducibility in research.

Exempting older research from the requirements of the proposed rule

It is imperative that research initiated prior to the effective date of this rule be exempt from the requirements of the rule. Excluding existing studies because the data cannot be shared would be a terrible threat to public health. What's more, many foundational human environmental public health studies cannot be replicated because people are no longer exposed to the hazards at high levels—often thanks to public health action taken in response to these same studies. Replicating the high levels of exposure observed in these studies is unethical and impossible, because it would involve intentionally exposing people to a harmful chemical. Even if these studies could be repeated, doing so would be a tremendous waste of time and funds.

The Supplemental Notice names technological barriers as one reason studies may not be able to comply with the requirements of the proposed rule. However, an equally important barrier is maintaining pre-existing pledges to protect participant privacy. A participant's right to privacy is established during informed consent to participate in a study. Informed consent is an ethical and legal requirement of human subjects research. Consent documents describe the nature of the research, the risks and benefits associated with participation, and how the researcher will protect the confidentiality of the research records. Up to the present, nearly all human environmental health studies have guaranteed in their informed consent not to share identifiable study data outside the research team.

As described earlier, we consider all human subjects data to be potentially identifiable, even after overt identifiers are removed. Thus, researchers will be put in the position of having to decide whether placing data in a restricted-access enclave—subject to access requirements that are not yet specified—violates their pre-existing pledge to participants. Rather than pressure researchers into sharing their data in ways that their informed consent could not have anticipated, potentially exposing them to legal liability, EPA should exempt all research initiated prior to the effective date of the Final rule from the requirements of the rule. To ensure the exemption is applied without bias, it should be applied on a blanket basis to all studies initiated prior to the passage of the rule—and not on a case-by-case basis by the administrator. The exemption should continue in perpetuity, and exempted research should carry equal weight to research that complies with the requirements of the Final rule.

Impacts of the Supplemental Notice on future participation in environmental health research

Should this rule become law, researchers will be able to plan future studies around the new requirements and obtain appropriate consent from participants at the outset. One alternative under consideration is broad consent, which permits wide–but not unlimited—sharing of identifiable data, including with investigators at other institutions and for future, unspecified research uses, without obtaining additional consent from the participant.

Some people may be comfortable with permissive data sharing models such as broad or open consent. For example, participants in the Personal Genome Project consented to share their data online with no guarantees of privacy (9). However, requesting consent for future data sharing at the study outset could negatively impact participation of racial and ethnic minorities, populations that are already underrepresented in health research (10), and over-burdened by diseases with known environmental triggers, such as asthma (11). For example, across multiple studies, African Americans have shown significantly less acceptance of or preference for broad consent models than White participants (12-14).

In addition, we recently published results from a survey investigating willingness to share environmental health data in adult women (15). The survey queried participants about their responses to vignettes describing study scenarios representing major environmental health research designs. Across all scenarios, public data sharing was a disincentive for 26% to 53% of participants (15). Certain types of data were perceived as more sensitive. Participants expressed less interest in participating in studies that accessed their electronic medical record (regardless of level of data sharing), and participants were more concerned about data sharing practices in studies of children: a greater proportion of participants reported that storing data for future research negatively impacted their decision to participate when the data were collected in children (28%–32%) compared to adults (4%–15%) (15). Nearly a quarter of all participants raised concerns about the privacy, security, and potential misuse of data stored for future research, including the potential for data breaches and re-identification (15). These results show that by opening access to research data, the proposed rule will likely dampen participation in future studies, particularly in vital studies of the effects of chemicals on children.

Ultimately, researchers will be responsible for communicating the implications of the proposed rule to their participants, by clearly defining what data will be shared, what security measures will be in place to protect the data, and what actions will be taken in the event of a breach, and discussing any potential re-identification risks that could result from their participation. The decisions that EPA makes now will have tremendous impact on people's confidence in participating in crucial public health research. We must not take advantage of their generosity with rules that threaten their privacy and discourage future participation in research.

Silent Spring Institute's foremost concerns are that EPA is able to use the best-available science to support regulatory decision-making without infringing on participants' rights to privacy. The Supplemental Notice of Proposed Rule falls short on both counts: it holds the potential to diminish the impact of irreplaceable public health research and suppress future research participation, and it proposes inadequate protections for participant privacy. The passage of this rule would represent a major failure by EPA to protect the health of all Americans.

Thank you for your consideration of these comments.

Sincerely,

Katherine E. Boronow

Julia Green Brody, Ph.D.

Silent Spring Institute

Attachments:

Boronow KE, Perovich LJ, Sweeney L, Yoo JS, Rudel RA, Brown P, et al. 2020. Privacy risks of sharing data from environmental health studies. Environmental Health Perspectives 128(1):017008.

Udesky JO, Boronow KE, Brown P, Perovich LJ, Brody JG. 2020. Perceived risks, benefits, and interest in participating in environmental health studies that share personal exposure data: A U.S. Survey of prospective participants. Journal of Empirical Research on Human Research Ethics 0(0):1556264620903595.

References:

1. Boronow KE, Perovich LJ, Sweeney L, Yoo JS, Rudel RA, Brown P, et al. Privacy Risks of Sharing Data from Environmental Health Studies. Environmental Health Perspectives. 2020;128(1):017008.

2. Sweeney L, Yoo J, Perovich L, Boronow K, Brown P, Brody J. Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. Technology Science. 2017;2017082801.

3. Brody JG, Morello-Frosch R, Zota A, Brown P, Perez C, Rudel RA. Linking exposure assessment science with policy objectives for environmental justice and breast cancer advocacy: the northern California household exposure study. Am J Public Health. 2009;99 Suppl 3:S600-9.

4. de Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD. Unique in the crowd: The privacy bounds of human mobility. Scientific Reports. 2013;3:1376.

5. Douriez M, Doraiswamy H, Freire J, Silva CT, editors. Anonymizing NYC taxi data: Does it matter? 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA); 2016: IEEE.

6. Siddle J. I Know Where You Were Last Summer: London's public bike data is telling everyone where you've been London. 2014 [Available from: https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html.

7. Freeman LB, Blair A, Hofmann J, Sandler DP, Parks CG, Thomas K. Agricultural Health Study Policy 2-4: Guidelines for Collaboration. 2017 [Available from: https://aghealth.nih.gov/collaboration/AHS%20Policy%202-4%20Guidelines%20for%20Collaboration_2017.1.pdf.

8. Sister Study. The Sister Study Data Sharing Policy. nd [Available from: https://www.sisterstudystars.org/Public/Sister/Documents/Data%20Access%20Policies%20and%20Procedures.pdf.

9. Zarate OA, Brody JG, Brown P, Ramirez-Andreotta MD, Perovich L, Matz J. Balancing Benefits and Risks of Immortal Data: Participants' Views of Open Consent in the Personal Genome Project. Hastings Cent Rep. 2016;46(1):36-45.

10. Konkel L. Racial and Ethnic Disparities in Research Studies: The Challenge of Creating More Diverse Cohorts. Environ Health Perspect. 2015;123(12):A297-302.

11. Forno E, Celedon JC. Health disparities in asthma. Am J Respir Crit Care Med. 2012;185(10):1033-5.

12. Ewing AT, Erby LA, Bollinger J, Tetteyfio E, Ricks-Santi LJ, Kaufman D. Demographic differences in willingness to provide broad and narrow consent for biobank research. Biopreserv Biobank. 2015;13(2):98-106.

13. Platt J, Bollinger J, Dvoskin R, Kardia SL, Kaufman D. Public preferences regarding informed consent models for participation in population-based genomic research. Genet Med. 2014;16(1):11-8.

14. Sanderson SC, Brothers KB, Mercaldo ND, Clayton EW, Antommaria AHM, Aufox SA, et al. Public Attitudes toward Consent and Data Sharing in Biobank Research: A Large Multi-site Experimental Survey in the US. Am J Hum Genet. 2017;100(3):414-27.

15. Udesky JO, Boronow KE, Brown P, Perovich LJ, Brody JG. Perceived Risks, Benefits, and Interest in Participating in Environmental Health Studies That Share Personal Exposure Data: A U.S. Survey of Prospective Participants. Journal of Empirical Research on Human Research Ethics. 2020;0(0):1556264620903595.