Preliminary Comments from Members of the Science Advisory Board

Mechanisms for secure access to personally identifying information (PII) and confidential business information (CBI) as discussed in the proposed rule consistent with existing laws and policies that protect PII and CBI.

(as of 8/21/19)

DR. HUGH A. BARTON 2

DR. DEBORAH BENNETT 3

DR. JANICE CHAMBERS 6

DR. SAMUEL COHEN 7

DR. JOHN GUCKENHEIMER 8

DR. KIMBERLY WHITE 9

Dr. Hugh A. Barton

1. Other agencies (e.g., NIH and HHS) use a tiered approach for access to PII data. Please comment on whether such an approach would be a good model for EPA to apply.

I cannot comment on the approaches used by other agencies as I have no experience with them.

2. Given the laws protecting CBI and PII, as well as the proposed requirements for data availability in the Strengthening Transparency in Regulatory Science proposed rule, please comment on how EPA could use studies involving CBI and/or PII to make regulatory decisions.

EPA needs to be able to use the best available scientific data and analyses. Ideally, these would be available either publicly or through access methods such as those referred to the previous question. Having said that, epidemiological studies in the past were done with the consent agreements that were used at the time. I am concerned that "Strengthening Transparency" will unreasonably apply a contemporary standard to studies done in the past and inappropriately make important information unavailable to EPA for regulatory purposes. I am aware of cases where pharmaceutical studies have been substantially impacted by the consent agreements used, leading to changes in agreement wording for future studies. There likely needs to be a similar differentiation of historical studies from future studies for EPA purposes as well.

It is unclear to me why toxicity studies in animals should ever be CBI, except to limit use of their data by other companies. It would seem that it should be possible to have the complete study publicly available (except perhaps proprietary information about a mixture or product as opposed to a single chemical), but require that submissions by other companies must obtain permission from the owners of the data/studies for which they might have to pay if they wish to use that data to support their regulatory needs.

Dr. Deborah Bennett

1. Other agencies (e.g., NIH and HHS) use a tiered approach for access to PII data. Please comment on whether such an approach would be a good model for EPA to apply.

The tiered approach should work very well for some types of studies, for example, studies involving exposures that were measured using biomarkers. In this case, one would want to be careful to limit the available dataset to include only those outcomes, exposures, and covariates that were used in the model to minimize the chance for de-identification.

One complexity may arise for studies with biomarkers of exposure if the study population was recruited from a very small region, such as a study recruiting woman delivering in a particular hospital. The HIPPA website noted in the background materials references provided by the EPA indicates for example, data should only be reported based on the geographical unit defined by the first 3-digits of a zip code. And, further, if the first three digits of a zip –code define a region less than 20,000 people, the data cannot be reported. In the case of, for example, babies born in a particular hospital, birth month would not provide sufficient de-identification, and quite possibly birth year would not provide enough de-identification. Great care would need to be taken with any study using a small recruiting base.

However, this question becomes more complicated if the exposures are based on the address, such as air pollution or pesticide exposure through databases such as the California Pesticide Use Registry, or similar state registries.

The web-site for the HIPPA Guidance document provided in the background section states: "A tiered approach provides access to research data using different strategies based upon disclosure risk. Access to information and data varies by tier. The greatest amount of information is available when access to data are most restricted. Replicating findings requiring PII information (e.g., residence) may not be possible with unrestricted public access. The amount of information available for analysis is dictated by the tier chosen." This statement recognizes that things like residences, and therefore by extension, residence based pollution measures, since those have been shown to be re-identifiable, would not need to be listed, hence protecting privacy.

But then, going back to the proposed transparency rule, it is stated that "Information is considered 'publicly available in a manner sufficient for independent validation' when it includes the 'information necessary for the public to understand, assess, and replicate findings.'" If spatially derived exposure levels are removed, the public can no longer replicate findings. It is not clear how the EPA would handle these apparently contradictory statements.

It is important to note there is a real concern regarding de-identification of data. The U.S. National Academy of Sciences, in their report *Improving Access to and Confidentiality of*

Research Data: Report of a Workshop, stated "Since unrestricted access can cause harm to individuals and also conflicts directly with respect for individual autonomy, it is not an appropriate policy (NRC 2000)."

Recently, a peer reviewed study examined the identifiability of records from an environmental health study in Northern California. Using data meeting HIPAA requirements to be de-identified, researchers were able to correctly identify over 25% of the participants (Sweeney et al. 2017). Another study searched the Lexis-Nexis database for stories that mentioned hospitalization, and used age, race, sex and Zip code from anonymized hospital admissions data base to match 43% of the people named in the news stories to their medical records (Sweeney 2015).

Consider Medicare information on deaths. If one were to consider exposure information assigned by zip code, the average number of deaths per year in a zip code is only 23. If one knew the age, race,
and sex, age at entry into Medicare (people who work after 65 often enter at a later age), year of death, and zip code of those people, it would not be difficult to identify them. This is why the HIPPA guidelines require information given by zip code to only reflect the first three digits of the zip code. But if exposures are assigned by zip code, even if the zip codes are not given, people are grouped into zip-code based groups. However, it is still possible to determine the zip code from the information that might be used in a study. Consider that you knew region of the country, an often used covariate. There are public maps showing how air pollution concentrations vary across the US, improving ones' ability to match a zip code to a location, especially if one considers the trends in year to year levels. Consider other potential covariates such as percent of the zip code that is age 65 or older, is black, is Hispanic, or is living below the poverty level, or items such as median house value in the zip code or the annual average temperature in the zip code. These could all be used to identify the zip code, from which then the individuals can be identified.

2. Given the laws protecting CBI and PII, as well as the proposed requirements for data availability in the Strengthening Transparency in Regulatory Science proposed rule, please comment on how EPA could use studies involving CBI and/or PII to make regulatory decisions.

The scientific community has developed several protocols and guidelines specifically designed for reporting study data, such as CONSORT, ARRIVE and STROBE. These protocols do not require public access to all study data but still improve the scientific basis of evaluating studies, and could be adopted by the EPA.

Death and birth certificate information is publically available from organizations such as state departments of health or the National Death Index, and hospital admissions data is available

from Medicare. To access this data, researchers sign Data Use Agreements prohibiting them from making public anything other than aggregate data summarizing statistics from large numbers of people.
Other researchers can, and have applied to the same organizations to obtain their own copies of the data, after signing their own Data Use Agreements. This approach could be considered an alternative for some research studies.

Sweeney L, Yoo JS, Perovich L, Boronow KE, Brown P and JG B. Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. Technology Science. 2017.

Sweeney L. Only You, Your Doctor, and Many Others May Know. Technology Science. 2015.

National Research Council. 2000. *Improving Access to and Confidentiality of Research Data: Report of a Workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9958.

Dr. Janice Chambers

Charge Questions

3. **Other agencies (e.g., NIH and HHS) use a tiered approach for access to PII data. Please comment on whether such an approach would be a good model for EPA to apply.**

The tiered approach seems to be a reasonable model for EPA to apply. For one thing, it is already used by at least other federal agencies, so it has a precedent for use by agencies that have concerns about access to PII of study participants. This type of approach could use the most restrictive tier initially for the greatest participant protection if this tier was sufficient to allow verification of the data sets or calculations in question. Only when the accessed information was insufficient for verification would there be a need to release more PII and thereby increase the risk for identification of the participants or their personal information.

4. **Given the laws protecting CBI and PII, as well as the proposed requirements for data availability in the Strengthening Transparency in Regulatory Science proposed rule, please comment on how EPA could use studies involving CBI and/or PII to make regulatory decisions.**

There are EPA laws protecting both CBI and PII and certainly there are Non-disclosure Agreements (NDA's) that are used in many situations to protect CBI, so there are mechanisms for this protection that have legal standing. EPA may need to obtain the protected information to make regulatory decisions and EPA provides assurance that its staff will not divulge protected information; this should not change in the future. EPA will need to provide the same assurances of protection for any individuals in the public who are granted permission to access the protected information, and these members of the public need to be fully informed about the legality of maintaining the confidentiality and the legal ramifications if the protection is violated.

Dr. Samuel Cohen

Dr. Cohen provided the following article:

Perignon, C., K. Gadouche, C. Hurlin, R. Silberman, and E. Debonnel. 2019. Certify reproducibility with confidential data. *Science* 12 Jul 2019, Vol. 365, Issue 6449, pp. 127-128. DOI: 10.1126/science.aaw2825
https://science.sciencemag.org/content/365/6449/127

Dr. John Guckenheimer

1. Other agencies (e.g., NIH and HHS) use a tiered approach for access to PII data. Please comment on whether such an approach would be a good model for EPA to apply.

I think that a tiered approach to PII would be a good model for the EPA to adopt. If possible, adopting the procedures used by the National Center for Health Statistics would simplify matters for everyone with an interest in this issue.

2. Given the laws protecting CBI and PII, as well as the proposed requirements for data availability in the Strengthening Transparency in Regulatory Science proposed rule, please comment on how EPA could use studies involving CBI and/or PII to make regulatory decisions.

EPA, alone or in partnership with other Federal agencies, should assemble databases containing the data about dose responses underlying its regulations. In the case of PII and CBI, public access to these data should be limited to those parties tasked with peer review and validation. From a long term perspective, EPA seeks to establish standards for experimental methodology and the modeling that connects the data to regulations. The public interest can be protected by policies that ensure that reasonable questions about data validity and analysis will be investigated by independent experts, similar to scientific peer review. A key aspect of making this process work is that EPA maintain a staff of experts who can manage access to confidential information.

John Guckenheimer

Dr. Kimberly White

Responses to the Questions

5. Other agencies (e.g., NIH and HHS) use a tiered approach for access to PII data. Please comment on whether such an approach would be a good model for EPA to apply.

   Response: Implementing a tiered approach as utilized currently by other federal agencies is a reasonable path-forward for EPA. The National Institutes of Health has extensive experience associated with the identification and handling of sensitive information. They have established guidance and policies to facilitate the process which can be evaluated and implemented, as relevant and appropriate by the EPA.

6. Given the laws protecting CBI and PII, as well as the proposed requirements for data availability in the Strengthening Transparency in Regulatory Science proposed rule, please comment on how EPA could use studies involving CBI and/or PII to make regulatory decisions.

   Response: Studies that may include confidential business information (CBI) and/or personally identifiable information (PII) can provide important information about exposure and human health impacts which may be critical for the regulatory process. In compliance with all appropriate laws protecting sensitive information, CBI and PII, EPA should utilize all relevant information at its disposable to make science-based regulatory decisions. Specifically, EPA should develop clear guidance and policies for its use of studies containing this type of information and how, when necessary and feasible, it will seek to make information available.