This comment responds to the Environmental Protection Agency's proposal to exclude from consideration in setting environmental standards studies where the raw, individual level data are not publicly available. I believe this proposal is inconsistent with the Clean Air Act mandate to base regulations on the "best available science" since almost all the best available science would be precluded from consideration by this proposal. It also conflicts with other Federal Government regulations to protect privacy.

Moreover, the gold standard of science is not re-analysis, but replication by other investigators in other populations. Replication is not the re-analysis of previously published studies. It occurs when other researchers independently conduct similar research and come to similar conclusions. In earlier standard setting for air pollution, for example, EPA based its conclusions on the r*eplicability* of the original estimates by these other studies. In the case of the PM2.5-mortality studies, a recent meta-analysis found 53 cohorts including in the U.S., Canada, Europe, and Asia[1]. Some of those cohorts were studied more than once, with updated analyses after more years of follow-up were obtained. Hence the results have been replicated many times by many groups in many countries. The same is true for ozone. Hence the burdens this regulation imposes on privacy is not necessary to assure the public that the science underlying the regulations has been replicated.

Human studies at academic institutions in the U.S. are regulated by HIPAA and by IRBs, as well as basic regards for privacy. In Europe, the are regulated by the General Data Protection Regulation. Other countries have similar requirements. This proposal conflicts with those regulations. Great progress in understanding the effects of pollution has been made by adding exposures to large cohort studies that were established to study cardiovascular disease and cancer. This includes analysis of metals in blood, urine, or toenails, assignment of air pollution exposure based on

residential address, etc. Precisely because these studies have measurements of many potential confounding factors, and assign exposure and other covariates based on address, it is difficult to make the data public without also making the participants identifiable. While some progress has been made in de-identifying data, studies of environmental exposures present more serious issues because often exposure is assigned based on geocodes, and neighborhood covariates are based on public geocoded data. This makes identifiability much easier. For example, our studies typically use neighborhood level variables such as population density, median household income, green space, percent of the population living in poverty, percent black, percent Hispanic, and local temperature. Given all this data, it is hard to deidentify the neighborhood. To illustrate the problem, following Hurricane Katrina, a local newspaper published a map of mortality locations. No roads were shown on the map and the only geographical data was neighborhood. Yet researchers were able to correctly identify the original residence for most of the locations in the published map[2].

A cohort study of pollution need not make public geocodes to endanger privacy if the underlying data is made public. It typically controls for 15-20 potential cofounders, usually including census-based measures of socio-economic status and other geocoded information. If those covariates were all dichotomous, this would result in over 32,000 unique combinations. If some of those variables are based on publicly available geocoded data, such as census tract measures of race, SES, population density, housing value, and county level data from CDC's Behavioral Risk Factor Surveillance survey, local air pollution levels, etc. it is possible to identify census tract of the participant. Since the number of people who die in a census tract in a year is small, if one knows the age, race, and sex of that person, which must be included in a data base to permit re-analysis of data, one knows who they are. Similar considerations apply to hospital admissions. With continuous confounders,

the situation is worse, since instead of only two values for each variable there are many. This issue has been well recognized, including by the National Academy of Sciences. In a report on the issue they stated "*In an experiment to discover whether confidentiality could be preserved while opening the data for public review, the study investigators attempted to disguise the identity of the study participants. They deleted as many features as possible from the questionnaires, such as the name, the state file number, the mother's maiden name, and the name of the person providing the information. However, they needed to retain a minimum set of features if other scientists were to be able to replicate the basic findings of the study…They found that even this minimum set of features could allow for identification of research participants.*[3]"

Recently, a study examined the identifiability of records from an environmental health study in Northern California. Using data considered by HIPAA to be sufficiently de-identified to be made public, they were able to correctly identify over 25% of the participants[4]. Another study searched the Lexis-Nexis database for stories that mentioned hospitalization, and by matching that with age, race, sex and Zip code from a supposedly anonymized hospital admissions data base was able to match 43% of the people named in the news stories to their medical records[5]. Since many obituaries are printed every day, and death certificates are publicly available the identifiability problem is even larger.

This can be illustrated with the Harvard Six City study. Participants were recruited from one neighborhood in each city, including Watertown MA, population 35,000. The average number of deaths per year in Watertown is 208. This is less than one death per day. Obviously knowing the date of death would uniquely identify most participants. But even if the data made public only included year of death, age, race, sex, and cause of death, with only 200 deaths in a year, most people can be

identified from those facts, before one even gets to other covariates included in the model.

The Canadian Community Health Survey Cohort followed 300,000 people and examined the association of $PM_{2.5}$ with mortality[6]. Because of privacy laws, the data was not given to the investigators, and analysis was performed on the computers at Statistics Canada. Yet this study is critical for EPA to consider as it reviews the adequacy of its current 12 $\mu g/m^3$ $PM_{2.5}$ standard, because essentially all the participants lived in locations below that standard. They found no threshold in the effects of PM2.5 in a population averaging 6.3 $\mu g/m^3$. How can EPA review that standard without considering it?

This rule will exclude European and Canadian studies of human participants from being considered by EPA in regulating environmental pollutants. The EU General Data Protection Regulation defines private data to included medical, physical, physiological, genetic, mental, economic, cultural or social identity of a person. Such data must be controlled by a data controller who must demonstrate that *any* use of the data has been consented to by the individual whose data it is. This obviously precludes making it publicly available.

EPA has argued that the data can be sufficiently de-identified to be made public, while still meeting their goal of allowing reanalysis. However, as is clear above, the number of variables included in the original analyses that would have to be omitted or condensed into crude categories is so large it would preclude any re-analysis from reproducing the original results. More plausible is EPA's argument that protected data centers could house the data and allow people to analyze it. But if the Canadian government would not allow the initial investigators to have the data mentioned above, it is unlikely that they would agree to convey it to an EPA computer, even with restricted access. This is likely true of most of the cohort

studies EPA relies upon, where IRBs or EU privacy rules are unlikely to allow transfer data to EPA or other US government computer centers.

As noted above, the gold standard of science is not re-analysis, but replication. In the case of the PM2.5-mortality studies the results have been replicated *many* times by *many* groups in *many* countries[7]. What value then has a re-analysis of an abbreviated subset of covariates from any given study? Particularly if it can't control for important covariates?

It is difficult to believe that EPA does not know that few human cohorts could comply with their requirements, and therefore difficult not to conclude that the real purpose of the proposal is to eliminate a vast body of highly relevant human data from being considered by EPA, resulting in a weakening or failure to tighten standards that no longer have "sufficient scientific evidence" by EPA's new criteria. Moreover, the argument of transparency is a charade. The general public is not going to download data or access computers to reanalyze this data, and EPA's proposal is really intended to allow industry groups, with clear financial conflicts of interest in the results of such studies to reanalyze the data until they can find a way to confuse the results. This is similar to the smoke and mirrors approach of the Tobacco industry regarding studies of second-hand smoke exposure, or when the lead industry redefined gasoline lead exposure in NHANES to make exposure 456 times higher in Washington DC than in Houston TX, and its association with blood lead go away.

This rule could allow interested parties to utilize the basic demographic and environmental data that are required for the analysis of most environmental epidemiology studies and identify and even contact study subjects who had been promised confidentiality by study investigators. This is not just a theoretical consideration. During the 1970's an industry group obtained the addresses of

participants in a US EPA study of air pollution (part of the CHESS studies) and contacted them to "verify the accuracy of the data they had provided to EPA." The participants were very unhappy with this, and EPA was forced to terminate the study. Terminating studies because of privacy concerns is not the way to obtain the best available science the Clean Air Act requires.

Nor is this just a privacy issue. In the context of epidemiology studies, a great deal of sensitive data must often be collected, data that could be damaging to individuals if disclosed by a third party. Moreover, none of this violation of individual rights would even address the basic scientific question: Have the results been replicated?

In exchange for excluding most cohort studies and threatening privacy, what would be gained by this approach. Recall that in EPA's previous NAAQS ISA reviews, a weight of the evidence approach was taken that asked whether these epidemiology studies reporting air pollution increased death rates, mostly from cardiovascular disease, were not just replicated, but were supported by other studies. For example, if air pollution was truly causing increased deaths from cardiovascular disease, one would expect that it also was associated with worsening of intermediary measures of health, such as blood pressure, cholesterol levels, atherosclerosis, inflammation, etc. that are established risk factors for cardiovascular deaths. Contrawise, if PM2.5 is increasing atherosclerosis, how could it not increase the risk of cardiovascular deaths? The recent ISA for PM2.5 reviewed the evidence and found just such supporting evidence. EPA has traditionally relied on supporting evidence such as this, along with true replication, to justify conclusions about health effects of pollutants.

For example, a follow-up of over 6500 patients in Ohio undergoing cardiac examinations found that PM2.5 was associated with increased risk of severe atherosclerosis and of having a myocardial infarction (heart attack) in the next three

years.[8] A twenty year follow-up study showed that PM2.5 exposure increased the risk of death in post MI patients, and that increase was larger in frailer people[9]. The ESCAPE study in Europe followed over 100,000 people who were free of coronary events at intake for 11.5 years, and found that a 10 $\mu g/m^3$ increase in $PM_{2.5}$ was associated with a 26% increase in the risk of a heart attack on follow-up[10].

But there are also experimental studies that support the association of exposure to PM2.5 with health outcomes that strongly support the epidemiology from the cohort studies of mortality. Two hundred homemakers in Taiwan were randomly assigned to receive a functional particle filter in their house, or an identical filtering device but with the actual filter element removed. Six home visits were conducted per year, and blood pressure was measured. The participants with the real filters had lower blood pressure than those with the sham filters, causally demonstrating that particle exposure increases blood pressure[11].

Moreover, toxicology studies also support the conclusions of the mortality studies. Wellenius exposed dogs to either filtered air or concentrated air particles, followed by a temporary blockage of the coronary artery. The animals exposed to particles experienced greater ischemia than those exposed to filtered air.[12, 13] A study of mice exposed to particles (15 $\mu g/m^3$) from the outside air showed that the particle exposure leads to more atherosclerotic plaque and increased macrophages and tissue factor in the plaques, which reduce plaque stability and increase the risk of a heart attack.[14] Another mouse study documented that particle exposure increased oxidation of low-density lipoprotein (making it much more dangerous), increased the thickness of the arterial wall, and promoted plaque growth and instability.[15]

So why risk exposing personal medical history of participants in studies showing cardiovascular mortality is related to PM2.5, or excluding most of those studies from consideration in standard setting, when there are dozens of studies

replicating the results, dozens of studies showing the expected associations with intermediary health effects, and experimental studies showing that PM2.5 raises blood pressure and increases atherosclerosis? Given this, what benefit outweighs the risk to participants? And what benefit to society outweighs the risk to society of not receiving the public health benefits the Clean Air Act grants them because EPA ignores much of the science? Or that individuals will henceforth refuse to participate in such studies?

I have illustrated the point for PM2.5 but the same applies to all regulations where EPA uses epidemiology studies to help set standards. The well-established protocols of the scientific community, looking at replication of results in other studies of other populations by other researchers, consistency with other epidemiologic results of more intermediary health effects, coherency with controlled exposure studies and toxicology is how conclusions of causality are drawn. And the transparency of this process, with, in the case of the NAAQS, includes multiple public meetings including public comments and detailed and critical discussion of individual studies and their strengths and weaknesses, ISA documents open for review, etc. is what provides the general public with the assurances they need that standards are science-based. Making public underlying medical information of individuals will add nothing to this, will generate great risk to the participants, and when investigators are unable to comply, will result in EPA ignoring the "best available science" due to an entirely self-inflicted injury.

The current proposal raises the specter of "secret science," suggesting a conspiracy of many scientists from multiple universities in multiple countries to published false or misleading studies, and to prevent transparent discussion and rigorous evaluation of important research. But paranoia is not a science policy, it is a medical disorder.

Sincerely,

Joel Schwartz

Professor of Environmental Health and Epidemiology

Harvard TH Chan School of Public Health,

Director, Harvard Center for Risk Analysis

1. Vodonos, A., Y. Abu Awad, and J. Schwartz, *The concentration-response between long-term PM2.5 exposure and mortality; A meta-regression approach.* Environmental Research, 2018.
2. Curtis, A.J., J.W. Mills, and M. Leitner, *Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina.* International Journal of Health Geographics 2006, , 2006. **5**: p. 44.
3. Council, N.R., *Access to Research Data in the 21st Century: An Ongoing Dialogue Among Interested Parties: Report of a Workshop.* . 2002, Washington, D.C.: National Academy Press.
4. Sweeney L, et al., *Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study.* Technology Science, 2017.
5. Sweeney, L., *Only You, Your Doctor, and Many Others May Know.* Technology Science, 2015.
6. Pinault, L., et al., *Risk estimates of mortality attributed to low concentrations of ambient fine particulate matter in the Canadian community health survey cohort.* Environmental Health, 2016. **15**(1): p. 18.
7. Vodonos, A., Y. Abu Awad, and J. Schwartz, *The Concentration-Response between Long-term PM2.5 Exposure and Mortality; A meta-regression approach.* Environmental Research 2018.
8. Hartiala, J., et al., *Ambient Air Pollution Is Associated With the Severity of Coronary Atherosclerosis and Incident Myocardial Infarction in Patients Undergoing Elective Cardiac Evaluation.* J Am Heart Assoc, 2016. **5**(8).
9. Gerber, Y., et al., *Frailty status modifies the association between air pollution and post-myocardial infarction mortality: a 20-year follow-up study.* J. Am Coll Cardiol., 2014. **63**: p. 1698-99.
10. Cesaroni, G., et al., *Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project.* BMJ, 2014. **348**: p. f7412.

11. Chuang, H.-C., et al., *Long-term indoor air conditioner filtration and cardiovascular health: A randomized crossover intervention study. .* Environment international, 2017. **106**: p. 91-96.
12. Bartoli, C.R., et al., *Concentrated ambient particles alter myocardial blood flow during acute ischemia in conscious canines.* Environ Health Perspect, 2009. **117**(3): p. 333-7.
13. Wellenius, G.A., et al., *Inhalation of concentrated ambient air particles exacerbates myocardial ischemia in conscious dogs.* Environ Health Perspect, 2003. **111**(4): p. 402-8.
14. Sun, Q., et al., *Ambient air particulate matter exposure and tissue factor expression in atherosclerosis.* Inhal Toxicol, 2008. **20**(2): p. 127-37.
15. Soares, S.R., et al., *Air pollution and antibodies against modified lipoproteins are associated with atherosclerosis and vascular remodeling in hyperlipemic mice.* Atherosclerosis, 2009. **207**(2): p. 368-73.