WRITTEN STATEMENT FROM ROY N. GAMSE TO THE US EPA SAB
REGARDING "STRENGTHENING TRANSPARENCY IN REGULATORY SCIENCE"
AUGUST 27, 2019

Thank you for the opportunity to provide comments for your consideration as you respond to the charge from EPA on the proposed regulation on Strengthening Transparency in Regulatory Science.

I am Roy Gamse. I worked for EPA for 10 years during the Nixon, Ford, Carter, and Reagan Administrations. I was Deputy Assistant Administrator, Deputy Associate Administrator, and Acting Assistant Administrator from 1977 to 1981, which included responsibility for the regulation development process and for economic and statistical analysis at EPA.

SUMMARY

1. **This proposal** has been justified on the false premise that identities of individuals in health research studies can be reliably masked to protect their privacy as required by law. They cannot be. Hence, this proposal **would result in many relevant studies being disallowed, including many that provide the basis for current EPA standards** that must be reviewed under applicable legislation.
2. **If this is such a good idea, why is it proposed just for EPA** rather than as legislation or regulation applying to all health-regulating agencies? Imagine that research used for FDA drug approvals had to pass the same requirements. FDA drug approvals would grind to a halt, just as would EPA health-protective regulations.
3. **The single most relevant document for the SAB to consider in responding to the EPA charge is the Comments of the International Society for Environmental Epidemiology on EPA's proposed rule on Strengthening Transparency in Regulatory Science.** Since the ISEE submitted it previously to EPA, but the SAB members have not seen it, I am attaching it to this submission. **SAB members <u>must</u> read this document. The ISEE recommends that EPA withdraw the proposal** in

part because the "masking" of personal identities cannot be reliably done and still allow reanalysis of the research.

DISCUSSION

Your review of this proposed regulation is critical because, if adopted, it runs the risk of sidetracking EPA's development of health-protective regulations. If this self-regulation were in place throughout EPA's history in combination with today's advanced techniques for "big data" analysis, many of EPA's vital current health protective regulations could not have been adopted. If it were required for all EPA regulations in the future, many regulations needed to protect health could not be promulgated, which would be a tragedy, negatively affecting the health and longevity of many in the US. Most critical, health-based regulations which are required by legislation to undergo periodic review (e.g., Clean Air Act ambient air quality regulations) could be decimated as the studies used as a basis for setting them in the past would be rejected under this rule.

As the record shows and will show, privacy considerations would disallow the use of much data that provide the basis for EPA's health-protective regulations. That is especially the case for epidemiological studies which track or measure the health of groups of individuals with different levels of exposure to pollutants, such as the studies used as the basis for EPA's ambient air quality standards and hazardous pollution standards.

Some people will tell you that masking personal information can be done to hide the identity of individuals. That is nice in theory, and it may have been more practical in the 1970s and 80s when EPA regulation of health impacting pollution was in its early stages. But you should learn, if you have not already, that recent techniques of data manipulation in the era of "big data" analysis will now allow specific identification of many individuals who are participants in studies that provide the basis of EPA regulation.

I direct your attention first to Dr. Joel Schwartz's article in the New England Journal of Medicine entitled " 'Transparency' as Mask? The EPA's Proposed

Rule on Scientific Data," published August 29, 2018 [https://www.nejm.org/doi/full/10.1056/NEJMp1807751].
Dr. Schwartz, Professor at Environmental Epidemiology at the Harvard School of Public Health and a MacArthur Foundation "genius award" winner, spells out with examples how individuals can be identified even after deleting their names and other identifying information, unless so much information is deleted that the research results cannot be verified.

Besides his technical explanations of how analysis can reveal identities of people who have a right to anonymity, Dr. Schwartz cites a few examples of how current analytic techniques can defeat efforts to hide individual identities:

- How addresses of victims of Hurricane Katrina were identified without published information on names or addresses. [Curtis, Mills, Leitner, Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina, International Journal of Health Geographics, October 10, 2006; [https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-5-44]
- A National Academy of Sciences experiment that showed confidentiality could not be preserved even after all participant information not required to allow scientists to replicate a study's findings were deleted. [National Research Council, Access to research data in the 21st century: an ongoing dialogue among interested parties, National Academy Press, 2002; https://www.nap.edu/catalog/10302/access-to-research-data-in-the-21st-century-an-ongoing]
- An environmental health study in California using data considered under HIPAA to be sufficiently de-identified to be made public, yet more than 25% of participants were correctly identified. [Sweeney, Perovich, Boronow et al, Re-identification risks in HIPAA safe harbor data: a study of data from one environmental health study, Technology Science, August 28, 2017; https://techscience.org/a/2017082801/]

- A Canadian Community Health Survey in which most people could be identified from relevant facts excluding names and other key identifiers [Pinault, Tjepkema, Crouse, 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; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750218/]

Since drafting these comments, I have discovered an even more vital document: *"Comments of the International Society for Environmental Epidemiology on EOAs proposed rule on Strengthening Transparency in Regulatory Science." [http://www.youreventinfo.org/ISEE/Documents/ISEE_Comments_on_EPA-HQ-OA-2018-0259-0001FINAL_ISEE_submitted.pdf].* These comments were submitted to EPA but not to the Science Advisory Board, so I am forwarding them here for the convenience of the SAB members, including the newer ones. The ISEE members are the true experts in the field of environmental epidemiology including whether and how identities of research subjects can be protected. They conclude (see especially Section 3 [pages 7 – 11] that modern data science can now overcome efforts to redact personal data and still allow the research to be re-analyzed. They also have very critical comments on EPA's dose-response analysis proposal in Section 4. **The ISEE conclusion: "We urge EPA to withdraw the proposal."**

I urge the SAB members to read the ISEE submission. If you are not persuaded by the ISEE document itself, I urge you to hold a follow-on review directly with expert members of the ISEE (who cannot participate in this session due to a directly conflicting international meeting overseas).

The examples described by Dr. Schwartz, and others included in the ISEE comments, show that masking doesn't adequately hide identities of research participants. Even if in some instances it could, the fact that in many instances it cannot will have a chilling effect on participation in the studies that are needed as a basis for EPA regulations. Ask yourself, would you allow your 8-year-old to be a participant in a study of the effects of lead exposure on children's intelligence if there was a not insignificant probability that her identity could be unmasked despite the efforts to protect it? Of course not. So fielding usable studies would be increasingly difficult.

The result is that this seemingly well-intentioned proposal will either stifle the research needed by EPA or prevent its use by EPA. One suspects that presumably well-meaning EPA leaders have been misled or fooled by proponents of this disastrous proposal.

Dr. Schwartz states " 'the gold standard' of science is not reanalysis, but replication," a view echoed by the ISEE. The SAB should redirect EPA's focus in that direction.
****
There is one more issue to be confronted in this consideration. **Why is this self-regulation being proposed just for EPA and not for all agencies involved in health-protective regulation?** There is no reason whatsoever that the same logic would not apply to regulations promulgated by the Occupational Safety and Health Administration, the Consumer Product Safety Commission, the National Highway Traffic Safety Administration, the Food and Drug Administration, and others. Why would EPA be regulated in this manner and not any other regulatory agency? Shouldn't this be a matter of legislation or regulation applying to every regulatory agency?

Simple answer: the effects of this regulatory proposal would cut both ways. Imagine the Food and Drug Administration being forbidden to approve new drugs unless the research justifying their use were subject to these same regulatory prohibitions. The privacy protections imagined in this proposal would not protect patient identities any more than they would subjects of research considered by EPA. Drug approvals would likely grind to a halt. As would EPA regulations for air quality standards, hazardous materials exposures, pesticides exposures, drinking water concentrations, etc.

This proposed self-regulation would make no sense for the FDA and it makes no sense for EPA. **Your answer to EPA should be that there is no way to provide certainty that privacy protections would not be violated, so approving this approach would eliminate the best use of research to underlie EPA regulation. It should be rejected.**

Signed: Roy N. Gamse

Secretariat
Infinity Conference Group
(Contact: Doreen Albertson)
1035 Sterling Road
Herndon, VA 20170

Beate Ritz., M.D., Ph.D., *President*
Manolis Kogevinas, M.D., Ph.D., *Past President*
Sara Adar, Sc.D, *Secretary-Treasurer*

703.925.0178 (Voice) 703.925.9453 (Fax)
secretariat@iseepi.org (E-Mail)
www.iseepi.org (website)

Comments of the International Society for Environmental Epidemiology on EPA's proposed rule on Strengthening Transparency in Regulatory Science (EPA-HQ-OA-2018-0259-0001)

The International Society for Environmental Epidemiology (ISEE) is the society that represents researchers who study environmental causes of ill health, including ambient air pollutants subject to the National Ambient Air Quality Standards (NAAQS) promulgated by EPA, as well as its standards for heavy metals, pesticides, drinking water, and other environmental contaminants. As such our members have supplied a substantial part of the research that is the basis of those standards. We write in strong opposition to the proposed changes in which studies are to be considered in setting such standards.

ISEE thinks that the proposed rule would deprive policy makers of the real-world epidemiologic evidence, based on real exposures of real people, that has been and will continue to be vital for future revisions of the NAAQS, drinking water standard, pesticide standards, and other health based standards based on a false understanding of transparency and replication in science. This is direct contradiction to requirements of the Clean Air Act, for example, to consider the best available science, and would create an approach to the use of scientific evidence on the adverse health effects of air pollution by EPA that would be at odds with the World Health Organization, the Royal Society for Medicine, and other national review bodies. It would also put EPA in direct opposition to the standards for considering evidence of health effects of smoking, high cholesterol, etc. by the Centers for Disease Control, of the studies considered for issuing guidelines generated by the National Institutes of Health, and of the Food and Drug Administration reliance on non-publically available studies to approve drugs for use by people.

The ostensible purpose of this rule is to provide "transparency" via internet access to underlying individual data, but there is no value per se in having individual data on the internet if this results in the exclusion of the vast majority of available human data. Most individual human data are excluded by data privacy laws in the US and other parts of the world from being made public in the manner proposed. Meanwhile, other currently available methods make much of this data available to researchers willing to sign data use agreements precluding them as well from making individual data public. ISEE strongly supports these efforts to make research transparent and works to encourage the broad critical discussion

of scientific results. In our view the real purpose of assuring "transparency" should be to provide assurance that the science that EPA relies on to set standards is valid. To that end, EPA already has in place a detailed review process, that, together with the information already made public about the data and methods used in each study, allows EPA and its external science advisory committees, to reach that judgment. It has been argued that public data would allow reanalysis of data to ensure conclusions of the studies EPA relies on are valid. We agree that reanalysis has a role to play and we cite several examples below where re-analyses of key studies have been conducted without compromising the privacy of research participants in the fashion currently being proposed. However, although data reanalysis has a role to play, ultimately, the key determination of the consistency of scientific evidence comes from replication, not reanalysis. Replication is not the re-analysis of previously published studies. It occurs when other researchers, independently conduct similar research on different populations in different locations, and come to similar conclusions. Science reaches consensus when competing hypotheses have been tested in multiple independent replications with broadly consistent results, not when one data set has been reanalyzed many times.

We oppose this proposal because:

1. Data privacy laws prevent information from studies involving humans from being made public in the manner proposed. As a result EPA will exclude, for example, all of the epidemiology studies linking air pollution to serious outcomes such as death, heart attacks, and strokes, disinfection byproducts to birth outcomes, toxic exposure to cognitive function, etc. from their rulemaking. Moreover, much of this data is already available to other researchers, just not in the approach EPA has proposed.
2. The essential data, study protocols, recruitment criteria, measurement techniques, and statistical modeling methods needed to evaluate the quality of the evidence have all been made publically available, and peer reviewed. The existing EPA review process provides ample opportunity for people to challenge these studies if, based on those data, they find them wanting.
3. The arguments that data can be redacted to be non-identifiable are belied by modern data science showing how easy it is to identify people with even more limited information than would be required to be made public.
4. The language on what types of dose response studies should be considered is both very specific and narrow, and would exclude many informative and valid approaches used in the scientific literature.

5. The ostensible benefits of this proposed rule cannot justify the exclusion of the most relevant science from consideration in setting NAAQS, safe drinking water standards, pesticide standards, and other health protective standards.

We address these points in turn.

1. Data Privacy Laws Prohibit making public the data EPA proposes to make public.

Research on the health effects of environmental exposures in people, by its very nature, includes sensitive information on the medical, physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. As such, requirements under the Health Insurance Portability and Accountability Act (HIPAA) and the National Death Index place restrictions on sharing these data. Similarly, Institutional Review Boards that must review all proposed research by universities and other government funded research organizations require the protection of data from study participants. Even investigators who have obtained death or birth certificate information from state departments of health, or hospital admissions data from Medicare, all sign Data Use Agreements prohibiting them from making public anything other than aggregate data summarizing statistics from large numbers of people. These requirements clearly prevent the investigators from complying with EPA's rule, and hence excludes their research. And yet, in some cases such as with the National Death Index and the Medicare data, these are publically available data. 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.

European and Canadian privacy laws also reject the idea that personal information from participants in research studies could ever be made public. Indeed, Europe has just tightened its data privacy laws with the General Data Protection Regulation. That regulation limits movement of private data outside of the EU, which would prohibit the type of data access EPA proposes. It defines private data to included medical, physical, physiological, genetic, mental, economic, cultural or social identity of that natural person and it states that the data controller must "demonstrate that the data subject has consented to processing of his or her personal data."(article 7). If the data were made public, even (if possible) in unidentifiable ways, the data controller could still not demonstrate that it was only being used for the purposes, and by the people to which the participant in the study consented.

As an example of the highly confidential personal information included in these types of studies, consider the Canadian Community Health Survey Cohort, which followed a cohort of 300,000 people and looked at the association of $PM_{2.5}$ with mortality[1]. The participants were linked to their tax records to obtain individual level information on income, which was used as a variable in the analysis. The data

includes individual education, marital status, age, sex, immigrant status, minority status, weight, smoking, diet, alcohol consumption, and neighborhood level census data. Because of the highly confidential nature of the data, even the investigators did not have the data. They performed their analyses on the computers at Statistics Canada, where the data resides. Yet this study is a critical study for EPA to consider as it reviews the adequacy of its current 12 $\mu g/m^3$ $PM_{2.5}$ standard, because essentially all of the participants lived in locations below that standard (the 95th percentile of exposure was 11.3 $\mu g/m^3$). The Canadian Statistics Act prevents these data from being made public, yet how can EPA set standards based on the best available science without considering it?

Or consider the recent study of the Medicare cohort by Di and coworkers[2]. One key result came from a restricted analysis including only people residing in U.S. ZIP Codes with annual $PM_{2.5}$ concentration below the 12 $\mu g/m^3$ NAAQS. That analysis included 247,682,367 person-years of follow-up and 11,908,888 deaths. They found a 1.36% (95% Confidence Interval (CI) 1.31%, 1.41%) increase in mortality rate per 1 $\mu g/m^3$ increase in $PM_{2.5}$ below 12 $\mu g/m^3$. Yet their Data Use Agreement with Medicare prohibits them from providing these data to anyone. How can EPA set a standard based on the best available science when such a huge study, with hundreds of millions of observations in the range relevant for determining the adequacy of the standard is excluded from consideration? Particularly when anyone else can apply to the Centers for Medicare and Medicaid Services for access to the data?

Or consider the ESCAPE study[3]. This study combined 22 cohorts across Europe, estimated address specific concentrations of multiple air pollutants for each participant, and examined the association of air pollution with mortality in these participants. European data privacy laws would certainly prevent the data from these people being made public as EPA proposes requiring before considering the results in setting standards for air pollution.

National laws prevent the kind of publication of data EPA proposes and these privacy protections are not baseless. Indeed 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[4]."

2. The key information about these studies is already publically available.

A. The EPA proposal is not necessary to assure people that the studies have been reasonably conducted because the study protocols, recruitment criteria, measurement techniques, and statistical modeling

methods including lists of the adjustment variables have all been made publically available, and peer reviewed. Tables of descriptive statistics are also already public. Moreover, the peer review process of journals means that a group of experts unrelated to the study have already commented on, and requested changes in, methods that they find wanting.

For example, the ESCAPE study consisted of 22 cohorts for which the methodology was well documented in papers describing how the recruitment was done, who was eligible to be recruited, who was not eligible to be recruited, how many people were recruited, where the recruitment was done, what kind of data were collected on the people, including medical data and data on risk factors for disease, and they have all produced detailed summary statistics of those data, and their correlations. Separate papers describe in detail how the land use regression models that produced the exposure were fit, what the protocol was for collecting monitoring data, locating the monitors, what land use variables were used and where they were obtained from etc. The statistical methods for fitting the land use regressions, and what variables were used were similarly described in detail. The papers that matched these two sets of data to examine the association of air pollution with health also described which data were used, how they were used, and what statistical models were fit. This is a completely transparent process, without the need for privacy violations. The same is true for the other cohort studies that EPA has relied on. They are perfectly transparent as to the nature of the population studied, how the recruitment was done, what data was collected on the individuals, and how the analysis was done.

Moreover, the scientific community has already adopted several protocols and guidelines specifically designed to improve reporting and evaluation of studies and improve the quality and transparency of the interpretation of findings. These protocols and guidelines, such as CONSORT[1], ARRIVE[2] and STROBE[3], do not require public access to all study data but still improve the scientific basis of evaluating studies.

B. The EPA Science Advisory Board review process already provides assessment of the adequacy of the approaches taken in the studies considered in setting a standard. For example, consider the setting of air quality standards. The process involves a summary of the science by the EPA in the Integrated Science Assessment (ISA), which is reviewed by experts, and then provided to the external EPA Clean Air Science Advisory Committee (CASAC) for review. If, based on how the analysis was done, what confounders were controlled for, or other criteria, the reviewers of the ISA, members of CASAC, or the general public find a study wanting, they are free to raise that issue in public comments to EPA on the draft ISA, which EPA is required to respond to substantively. Subsequently they can again provide comments to CASAC when it reviews the draft ISA. This has happened routinely in the past, where

many comments on the adequacy of individual studies, on possible flaws in measurement techniques or analytical methods have been raised and debated in past NAAQS reviews under the current review process for setting the NAAQS. There is no need to exclude most human studies to achieve a goal that is already being met.

C. The replication process of science also provides protections against inappropriate methods or other study protocols. This is because dozens of studies have been done, in different ways, by different teams of investigators. For example, to examine the risk of mortality from ambient air pollution the ESCAPE study used a standard Cox Proportionate Hazard Model; the American Cancer Society (ACS) study, with different investigators, has used a modified Cox model with spatially varying effects[5]; a reanalysis of the Six City Study used a Poisson survival analysis with covariate effects allowed to vary every year[6]; an analysis of Massachusetts death data used a relative incidence logistic regression model where deaths from causes not related to air pollution served as controls[7], etc. This wide variability in statistical approaches, along with the transparency in methods and with the expert peer review of CASAC for air and other Science Advisory Board committees for pesticides and water (which can instruct EPA to de-emphasize studies found problematic) means that any reasonable goals of assurance that the standard is set based on valid science can be met with the current process.

It is sometimes argued that the failure to replicate studies is a “crises in science”, and that necessitates the EPA proposal. But the EPA proposal will only facilitate reanalysis, and compared to replication, that is a thin reed. Reanalysis can determine whether the same data, analyzed in a different manner (which may or may not be appropriate) produces similar results. Replication, in contrast, tests whether different populations recruited in different ways in different locations, with different data collected on them, or in a different manner, and potentially analyzed in a different manner, as in the examples above, produces similar results. This is why scientists believe that replication is the key approach to reaching conclusions.

We have already noted that administrative data, such as mortality and Medicare data in the U.S. is already publically available, but in ways that protect privacy, unlike the EPA proposal. In addition many cohort studies incorporate a process whereby outside investigators can apply to analyze different hypotheses in the cohort. Indeed, this is how most of the cohort studies of air pollution and death have occurred; with new investigators applying to do additional analyses in cohorts not originally set up to examine air pollution. This process, unlike the EPA proposal, maintains privacy for the participants of the studies.

Finally, re-analysis of some key studies is something that has already been done, without the damaging requirements for public access that EPA has proposed. For example, the Health Effects Institute, a joint venture between scientists, government, and industry, funded the reanalysis of two early air pollution cohorts by an independent team and found similar results under the supervision of an advisory board that included representatives of industry. The data from Needleman's key 1979 study of lead and IQ[8] were also reanalyzed by another investigator using different methods to address a number of criticisms, again, without making the data public[9]. Once replication of Needleman's studies results by many other studies that investigated lead cohorts was achieved, the issue of reanalysis, appropriately, faded away.

3. The arguments that data can be redacted to be non-identifiable are belied by modern data science

EPA argues that it would be straightforward to anonymize data so that it can be made public. Most of the agencies responsible for the data involved disagree. For example the center for Medicare and Medicaid services, which is part of the same government as EPA, prohibits people who obtain Medicare information for research purposes from making anything other than summary information across large numbers of people private or public. European privacy laws would clearly prohibit making public the data behind the cohort studies that have been done. And there is good reason or their concern. Consider the Harvard Six City study. To get good exposure estimates participants were recruited not across the cities involved but from one neighborhood in each city. In Boston that neighborhood was Watertown with a population of about 35,000. From data obtainable from the MA department of Public Health, the average number of deaths per year in Watertown is 208. Applying the average US mortality rate per 1000 people to the population of Watertown would give very similar results. This is less than one death per day. Obviously knowing the date of death alone would uniquely identify most of the participants. But even if the data made public were restricted to providing only the year of death (which is required to replicate the analysis, or do any reasonable analysis of the effect of air pollution on annual mortality) the analysis of such data also requires knowing the age, race, sex, and cause of death of each individual. Again with only 200 deaths in a year, it is likely that most people can be identified from knowing those facts. After all, that is only 100 deaths per year for each sex, and those deaths occur across a range of at least 40 ages at death, and many different causes. Combining this with publically available obituary data, or publically available death certificate data, would complete the identification of each person.

The Medicare cohort which followed over 60 million Medicare participants might seem less susceptible to this issue. But consider that exposure was assigned to people by Zip code. From the data provided in the paper, the average number of deaths per year in a Zip code was 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. Since all of that data except the Zip code would have to be made public under EPA's proposal, the question of whether the data are sufficiently de-identified becomes how difficult would it be to identify the Zip code of the participants, given all the information used in the analysis, which EPA would like to make public? Since everyone in a given Zip code was assigned the same exposure and same covariates in the same year, it is easy to group observations into Zip codes, even without their identity being known. After this, can they be identified? The data that was used in the analysis of the Medicare cohort included region of the country. So the matching is already made easier by knowing which of 10 regions each Zip code is in. Then, for each Zip code, the annual average of ozone and $PM_{2.5}$ is part of the analysis data. There is considerable variability in both of these variables, and, moreover, there are public maps showing how they vary across the US, including in the published paper of Di and coworkers. EPA also has ~1800 monitoring sites to provide guidance on what levels prevail where, in what year. Using those, it should be easy to assign each Zip code to a sub region that corresponds to that range of $PM_{2.5}$ and ozone. But over 13 years, these exposures changed differently over time in different Zip codes, and since annual analyses were done in the paper, that data would have to be available under the proposed EPA rules as well. Looking at levels and trends can assign the unknown Zip codes to much smaller sub regions. Then consider that the analysis controlled for, and therefore under EPA's proposal, should make public, percent of the Zip code that is age 65 or older, is black, is Hispanic, percent of people 65 and older living below the poverty level, or not having completed high school, percent of owner occupied housing in the Zip code, population density in the Zip code, median house value in the Zip code, percent of owner occupied housing in the Zip code, and the annual average temperature in the Zip code. All of these data are, of course publically available from the U.S. Census with the Zip codes *identified*. How hard is it then to match? Consider median housing value. This varies greatly in the US, and even between nearby neighborhoods. Restricted to a sub-region, this provides a lot of information on where a Zip code could possibly be. So does mean annual temperature, and all of the other variables listed. The analysis data includes additional variables that were controlled for, including county level percent smokers and mean body mass index, and percent of people age 65 and over who had hemoglobin A1c and LDL cholesterol tested each year, and percent who had an annual checkup, by hospital service area. All of this data is publically available with the location identified. How hard would it be to match? Indeed, consider the simple calculation that if each of the 15 area level variables were classified as high vs low, there would be 32,768 unique combinations in each region which contains only about 3900 Zip codes. And of course those variables are continuous, giving many more unique combinations. And this

is without the ozone and $PM_{2.5}$ concentrations. Given all this information, we believe that most and likely all of the individual Zip codes, and hence most of the individuals who died, would be identifiable. About a third of the Medicare participants died during follow-up in that study. This would certainly violate rules of the Center for Medicare and Medicaid Services for use of their data, as well as general prevailing standards of individual privacy.

The issue of identifiability of data based on limited public information has been known for some time. Famously, Latanya Sweeney as a student at MIT (now at Harvard) identified the personal hospital admission record of Massachusetts Governor William Weld based on his date of birth and Zip code, and publically available voting registration data that includes names, addresses, and birth dates. The Safe Harbor provision of HIPAA offers guidance to researchers by prescribing how to redact data for public sharing, which may be what EPA is referring to when it discusses known methods of de-identification. Recently, a peer reviewed 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, which involved far fewer variables than would be required to make public in the cohort studies, they were able to correctly identify over 25% of the participants[10]. 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[11]. Since large numbers of obituaries are printed every day, the numbers that could be identified by such a process is even larger. And that is before considering the other data available to help matching, such as the variables described above.

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*.[12]"

Identification of participants in studies, in addition to possible violations of the law, could have long term consequences for future studies. People will be much less likely to agree to participate in long term epidemiological studies if they have seen that people in prior studies have been identified. The Centers for Medicare and Medicaid Services would likely cut off access to all investigators to their data in the

future. Concern about privacy would not only prevent future research on environmental exposures, but also have widespread ramifications for health research in general, since many of the cohort studies of air pollution (e.g. Framingham Heart Study, American Cancer Society Study, Nurses Health Study, MESA study) were originally funded for other purposes and continue to do important research on cardiovascular disease and cancer.

While we have focused this discussion on air pollution studies, the same issues apply to studies of lead, mercury, arsenic, pesticides, PFOA and PFAS, water disinfection byproducts, nitrates in drinking water, etc. For most studies publishing the raw data that the EPA seeks in this proposal would violate national laws and thus will not happen, forcing EPA to ignore much of the relevant science.

Although for U.S. based studies, the laws are less clear but publishing identifiable data is generally prohibited, and as noted above, most of these studies have enough information on enough variables, which is required to analyze the data with proper control for potential confounders, that the data is effectively identifiable. This issue of identifiability prevents public access such as that proposed by EPA but it does not prevent interested investigators from filing their own requests for data to the Center for Medicare and Medicaid Services, or to the public health departments of states for the death or birth certificate data, or cancer and birth defect registries that has served as the basis for many environmental health studies examining air pollution, water pollution, heavy metals, toxic substances, and pesticides. Indeed hundreds of investigators have applied and independently obtained access to these data, and there are many publications by different investigators using that data for studies of air and water pollution, other than the original investigators. The key is that each has signed their own Data Use Agreement, pledging to keep the data secure, and not breach privacy.

This existing system has worked well, providing the EPA with studies of mortality and hospital admissions studies from the same U.S. data by multiple investigating teams, while still maintaining data privacy. The publication of similar studies, by other investigating teams, using equivalent data from other countries, also provides the EPA with the assurance it needs that the results they rely on have been replicated, and by different people using different approaches. For example, Medicare data was first used to look at air pollution by one investigator, first working at EPA and then at Harvard[13]. Then other investigators sought the data, and published their own studies[14-19]. So while these data may not be made publically available as EPA proposes, it is available to investigators with the will to apply for access and adhere to the data use rules. Similarly, many studies, by many different teams of investigators have used mortality data from state mortality or national records to examine the acute effects of air pollution[20-59]. There is no reason to risk revealing identified confidential medical information to accomplish a task that

is already accomplished. Moreover, EPA, or its science advisors have asked investigators to perform additional analyses of their cohort studies, or provide additional data, and the authors have generally been responsive. So there is no need to risk privacy violations to accomplish this task, either.

In summary, data privacy laws in most countries would prevent EPA from considering studies from those countries, based on legitimate concerns about both identifiability and consent. In the U.S. it is clear that the data EPA would require be made public would allow identification of people in violation of privacy standards, and with widespread damage to public health research if it were done. Therefore, this proposal would prevent EPA from considering essentially all of the literature relating long term exposure to air pollution to mortality, and to most other outcomes, as well as studies that investigated pesticides, water pollution etc.

4. The EPA's proposal in regards to dose-response analysis is inappropriate and not scientifically based

EPA's proposal details very specific criteria for judging what types of analyses provide valuable information about what concentration or dose-response curves look like, instead of relying on the judgment of the scientific community and its own science advisory committees. The proposal states "When available, EPA shall give explicit consideration to high quality studies that explore: a broad class of parametric dose-response or concentration-response models; a robust set of potential confounding variables; nonparametric models that incorporate fewer assumptions; various threshold models across the dose or exposure range; and models that investigate factors that might account for spatial heterogeneity." But this ignores studies that sequentially exclude observations with exposures (doses) above a certain level, although they can unambiguously demonstrate that an association exists below that level, and subject to power limitations, that it does not exist below that level. It also puts an inappropriate emphasis on studies looking at broad classes of parametric forms for dose response, which is an approach little used in environmental epidemiology, because those parametric forms by their nature can force dose-response curves into shapes not indicated by the data. Moreover, giving higher priority to a study because it tried a large number of mostly inappropriate parametric forms compared to another study that considered a smaller number of well thought out parametric forms creates major multiple comparison issues and stands science on its head. Multiple bad analyses do not make a study more reliable. Instead, environmental epidemiology prefers penalized splines and similar non-parametric approaches. We know of no epidemiology society that recommends EPA's approach.

The emphasis on threshold models is also questionable. A threshold model is a form of a spline, a model where for example, the slope of a linear association is allowed to change at a certain point. In a threshold

model, one considers a piecewise linear dose-response, where the slope is assumed to be zero below some dose T, and some positive number above T. But that is a very restrictive assumption. Spline models are quite common in environmental epidemiology, but rather than, in the example above, forcing the slope below T to be zero, they allow the slope to change at T, but allow the slope below T to take on any value the data dictates. EPA provides no reason to justify the approach they propose, or for constraining the slope to be zero when the data might differ. Moreover, when EPA states that there is growing evidence for nonlinear dose-response relationships they imply that science is finding more evidence of thresholds. In fact, in human studies the opposite is true. For lead, for example, the nonlinearity is a flattening out of the dose-response relationship at the high dose end, not at the low dose end. This is true for many other toxins as well, including $PM_{2.5}$. And there is a clear biological reason—most of these toxins act on pathways in the human body that produce the diseases of aging. These pathways have already had their coping abilities overwhelmed, and hence we age. Therefore, any incremental stimulus along these pathways would be expected to produce incremental damage to the body. It is for this reason that a National Academy of Sciences committee on risk assessment stated "The committee recommends that cancer and non-cancer responses be assumed to be linear as a default[60]."

The methods for assessing the shape of a concentration-response relationship is a scientific question, and better left to scientists and EPA's science advisor, without stating preferences with no scientific justification given, and no apparent input from either the scientific community or EPA's own scientific advisors.

5. **The EPA proposal would not provide any gains on the transparency of science, but instead would prevent EPA from setting standards based on the best available science.**

As noted above, essentially all non-U.S. studies would be legally unable to meet EPA's requirements. The Canadian Community Health Study data would remain safe behind protected servers at Statistics Canada. While the protocols, methods, and all other information necessary to judge the study are available, EPA would nevertheless ignore it. The same for the ESCAPE study, the UK national cohort study, the Netherlands National Cohort study, the Danish National Birth Cohort and the Norwegian Birth Cohort (both funded by NIH), etc. Similarly, U.S. cohorts, given a choice between violating their promises in the informed consent documents they signed or their Data Use Agreements, or not making their data public, will choose the latter. EPA will achieve no greater "transparency", they will only deny themselves the ability to set standards based on the studies that the rest of the world relies on. Indeed,

EPA's own Science Advisory Board has commented that this proposed rule" could have the effect of removing legal, ethical, and peer-reviewed studies of health effects as sources to support the agency's regulatory efforts". It would also put the EPA at odds with other government agencies, such as the Centers for Disease Control, which bases judgements and risks assessments on tobacco smoke on epidemiology studies whose individual data is not public in the way EPA proposes.

The only thing that would ostensibly be gained by EPA's rule would be, were investigators to comply with EPA's request, the ability for others to re-analyze the same data. As noted above, administrative Medicare and mortality data is already available to all investigators who apply and agree to data privacy conditions, many cohort studies provide access to outside investigators, and replication of cohort studies by other cohort studies provides the same assurance that the results are not due to the nature of the data or the analytical approach of one group of investigators.

The EPA's attempt to prescribe how scientist should assess the shape of concentration-response relationships again seems designed more to tie EPA's hands and prevent it from considering studies most scientists think are relevant, rather than to serve any public health purpose.

The major scientific issue challenging EPA's regulations is not the use of human studies that have been replicated multiple times. It is that EPA has to base many of its regulations on extrapolation from animal studies because human studies, particularly at relevant doses, have not been available. This proposal, by announcing that human studies, when done, will likely be ignored, will only exacerbate this problem. We urge EPA to withdraw the proposal.

References:

1. Pinault L, Tjepkema M, Crouse DL, Weichenthal S, van Donkelaar A, Martin RV, Brauer M, Chen H and Burnett RT. Risk estimates of mortality attributed to low concentrations of ambient fine particulate matter in the Canadian community health survey cohort. *Environmental Health*. 2016;15:18.

2. Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Choirat C, Dominici F and Schwartz JD. Air Pollution and Mortality in the Medicare Population. *N Engl J Med*. 2017;376:2513-2522.

3. Beelen R, Raaschou-Nielsen O, Stafoggia M, Andersen ZJ, Weinmayr G, Hoffmann B, Wolf K, Samoli E, Fischer P, Nieuwenhuijsen M, Vineis P, Xun WW, Katsouyanni K, Dimakopoulou K, Oudin A, Forsberg B, Modig L, Havulinna AS, Lanki T, Turunen A, Oftedal B, Nystad W, Nafstad P, De Faire U, Pedersen NL, Ostenson CG, Fratiglioni L, Penell J, Korek M, Pershagen G, Eriksen KT, Overvad K, Ellermann T, Eeftens M, Peeters PH, Meliefste K, Wang M, Bueno-de-Mesquita B, Sugiri D, Kramer U, Heinrich J, de Hoogh K, Key T, Peters A, Hampel R, Concin H, Nagel G, Ineichen A, Schaffner E, Probst-Hensch N, Kunzli N, Schindler C, Schikowski T, Adam M, Phuleria H, Vilier A, Clavel-Chapelon F, Declercq C, Grioni S, Krogh V, Tsai MY, Ricceri F, Sacerdote C, Galassi C, Migliore E, Ranzi A, Cesaroni G, Badaloni C, Forastiere F, Tamayo I, Amiano P, Dorronsoro M, Katsoulis M, Trichopoulou

A, Brunekreef B and Hoek G. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. *Lancet*. 2014;383:785-95.

4. Statistics CoN. *Improving Access to and Confidentiality of Research Data: Report of a Workshop.* , . washington,, D.C.: National Academy Press; 2000.

5. Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y, Turner MC, Pope CA, 3rd, Thurston G, Calle EE, Thun MJ, Beckerman B, DeLuca P, Finkelstein N, Ito K, Moore DK, Newbold KB, Ramsay T, Ross Z, Shin H and Tempalski B. Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. *Res Rep Health Eff Inst*. 2009:5-114; discussion 115-36.

6. Lepeule J, Laden F, Dockery D and Schwartz J. Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009. *Environ Health Perspect*. 2012;120:965-70.

7. Kloog I, Ridgway B, Koutrakis P, Coull BA and Schwartz JD. Long- and short-term exposure to PM2.5 and mortality: using novel exposure models. *Epidemiology*. 2013;24:555-61.

8. Needleman HL, Gunnoe C, Leviton A, Reed R, Peresie H, Maher C and Barrett P. Deficits in psychologic and classroom performance of children with elevated dentine lead levels. *N Engl J Med*. 1979;300:689-95.

9. Schwartz J. Beyond LOEL's, p values, and vote counting: methods for looking at the shapes and strengths of associations. *Neurotoxicology*. 1993;14:237-46.

10. 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.

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

12. Council NR. *Access to Research Data in the 21st Century: An Ongoing Dialogue Among Interested Parties: Report of a Workshop.* . Washington, D.C.: National Academy Press; 2002.

13. Schwartz J. PM10, ozone, and hospital admissions for the elderly in Minneapolis-St. Paul, Minnesota. *Arch Environ Health*. 1994;49:366-74.

14. Fuchs VR and Frank SR. Air pollution and medical care use by older Americans: a cross-area analysis. *Health Aff (Millwood)*. 2002;21:207-14.

15. Wellenius GA, Schwartz J and Mittleman MA. Air pollution and hospital admissions for ischemic and hemorrhagic stroke among medicare beneficiaries. *Stroke*. 2005;36:2549-53.

16. Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL and Samet JM. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Jama*. 2006;295:1127-34.

17. Suh HH, Zanobetti A, Schwartz J and Coull BA. Chemical properties of air pollutants and cause-specific hospital admissions among the elderly in Atlanta, Georgia. *Environ Health Perspect*. 2011;119:1421-8.

18. Anderson GB, Dominici F, Wang Y, McCormack MC, Bell ML and Peng RD. Heat-related emergency hospitalizations for respiratory diseases in the Medicare population. *Am J Respir Crit Care Med*. 2013;187:1098-103.

19. Bell ML, Ebisu K, Leaderer BP, Gent JF, Lee HJ, Koutrakis P, Wang Y, Dominici F and Peng RD. Associations of PM(2).(5) constituents and sources with hospital admissions: analysis of four counties in Connecticut and Massachusetts (USA) for persons >/= 65 years of age. *Environ Health Perspect*. 2014;122:138-44.

20. Lipfert FW. Air pollution and mortality: specification searches using SMSA-based data. *J Environ Econ Manage*. 1984;11:208-43.

21. Katsouyanni K, Karakatsani A, Messari I, Touloumi G, Hatzakis A, Kalandidi A and Trichopoulos D. Air pollution and cause specific mortality in Athens. *J Epidemiol Community Health*. 1990;44:321-4.

22. Ostro BD. Associations between morbidity and alternative measures of particulate matter. *Risk Anal*. 1990;10:421-7.

23. Fairley D. The relationship of daily mortality to suspended particulates in Santa Clara County, 1980-1986. *Environ Health Perspect*. 1990;89:159-68.

24. Schwartz J and Marcus A. Mortality and air pollution in London: a time series analysis. *Am J Epidemiol*. 1990;131:185-94.

25. Kinney P and Ozkaynak H. Associations of daily mortality and air pollution in Los Angeles County. *Environ Res*. 1991;54:99-120.

26. Schwartz J. Particulate air pollution and daily mortality in Detroit. *Environ Res*. 1991;56:204-13.

27. Pope CA, 3rd, Schwartz J and Ransom MR. Daily mortality and PM10 pollution in Utah Valley. *Arch Environ Health*. 1992;47:211-7.

28. Dockery DW, Schwartz J and Spengler JD. Air pollution and daily mortality: associations with particulates and acid aerosols. *Environ Res*. 1992;59:362-73.

29. Spix C, Heinrich J, Dockery D, Schwartz J, Volksch G, Schwinkowski K, Collen C and Wichmann HE. Air pollution and daily mortality in Erfurt, east Germany, 1980-1989. *Environ Health Perspect*. 1993;101:518-26.

30. Schwartz J. Air pollution and daily mortality in Birmingham, Alabama. *Am J Epidemiol*. 1993;137:1136-47.

31. Saldiva PH, Lichtenfels AJ, Paiva PS, Barone IA, Martins MA, Massad E, Pereira JC, Xavier VP, Singer JM and Bohm GM. Association between air pollution and mortality due to respiratory diseases in children in Sao Paulo, Brazil: a preliminary report. *Environ Res*. 1994;65:218-25.

32. Anderson HR, Limb ES, Bland JM, Ponce de Leon A, Strachan DP and Bower JS. Health effects of an air pollution episode in London, December 1991. *Thorax*. 1995;50:1188-93.

33. Ostro B. Fine particulate air pollution and mortality in two Southern California counties. *Environ Res*. 1995;70:98-104.

34. Samet JM, Zeger SL and Berhane K. The association of mortality and particulate air pollution. I: Particulate Air Pollutoon and Daily Mortality. The Phase I Report of the Particle Epidemiology Evaluation Project. . 1995.

35. Ballester F, Corella D, Perez-Hoyos S and Hervas A. Air pollution and mortality in Valencia, Spain: a study using the APHEA methodology. *J Epidemiol Community Health*. 1996;50:527-33.

36. Dab W, Medina S, Quenel P, Le Moullec Y, Le Tertre A, Thelot B, Monteil C, Lameloise P, Pirard P, Momas I, Ferry R and Festy B. Short term respiratory health effects of ambient air pollution: results of the APHEA project in Paris. *J Epidemiol Community Health*. 1996;50 Suppl 1:s42-6.

37. Sunyer J, Castellsague J, Saez M, Tobias A and Anto JM. Air pollution and mortality in Barcelona. *J Epidemiol Community Health*. 1996;50 Suppl 1:s76-80.

38. Wojtyniak B and Piekarski T. Short term effect of air pollution on mortality in Polish urban populations--what is different? *J Epidemiol Community Health*. 1996;50 Suppl 1:S36-41.

39. Zmirou D, Barumandzadeh T, Balducci F, Ritter P, Laham G and Ghilardi JP. Short term effects of air pollution on mortality in the city of Lyon, France, 1985-90. *J Epidemiol Community Health.* 1996;50 Suppl 1:S30-5.

40. Loomis DP, Borja-Aburto VH, Bangdiwala SI and Shy CM. Ozone exposure and daily mortality in Mexico City: a time-series analysis. *Res Rep Health Eff Inst.* 1996:1-37; discussion 39-45.

41. Kelsall JE, Samet JM, Zeger SL and Xu J. Air pollution and mortality in Philadelphia, 1974-1988. *Am J Epidemiol.* 1997;146:750-62.

42. Woodruff TJ, Grillo J and Schoendorf KC. The relationship between selected causes of postneonatal infant mortality and particulate air pollution in the United States. *Environ Health Perspect.* 1997;105:608-12.

43. Sartor F, Demuth C, Snacken R and Walckiers D. Mortality in the elderly and ambient ozone concentration during the hot summer, 1994, in Belgium. *Environ Res.* 1997;72:109-17.

44. Samet J, Zeger S, Kelsall J, Xu J and Kalkstein L. Does weather confound or modify the association of particulate air pollution with mortality? An analysis of the Philadelphia data, 1973-1980. *Environ Res.* 1998;77:9-19.

45. Pereira LA, Loomis D, Conceicao GM, Braga AL, Arcas RM, Kishi HS, Singer JM, Bohm GM and Saldiva PH. Association between air pollution and intrauterine mortality in Sao Paulo, Brazil. *Environ Health Perspect.* 1998;106:325-9.

46. Diaz J, Garcia R, Ribera P, Alberdi JC, Hernandez E, Pajares MS and Otero A. Modeling of air pollution and its relationship with mortality and morbidity in Madrid, Spain. *Int Arch Occup Environ Health.* 1999;72:366-76.

47. Lee JT, Shin D and Chung Y. Air pollution and daily mortality in Seoul and Ulsan, Korea. *Environ Health Perspect.* 1999;107:149-54.

48. Zeger SL, Dominici F and Samet J. Harvesting-resistant estimates of air pollution effects on mortality. *Epidemiology.* 1999;10:171-5.

49. Samet JM, Dominici F, Curriero FC, Coursac I and Zeger SL. Fine particulate air pollution and mortality in 20 U.S. cities, 1987-1994. *N Engl J Med.* 2000;343:1742-9.

50. Wichmann HE, Spix C, Tuch T, Wolke G, Peters A, Heinrich J, Kreyling WG and Heyder J. Daily mortality and fine and ultrafine particles in Erfurt, Germany part I: role of particle number and particle mass. *Res Rep Health Eff Inst.* 2000:5-86; discussion 87-94.

51. Goldberg MS, Burnett RT, Bailar JC, 3rd, Brook J, Bonvalot Y, Tamblyn R, Singh R and Valois MF. The association between daily mortality and ambient air particle pollution in Montreal, Quebec. 1. Nonaccidental mortality. *Environ Res.* 2001;86:12-25.

52. Roemer WH and van Wijnen JH. Daily mortality and air pollution along busy streets in Amsterdam, 1987-1998. *Epidemiology.* 2001;12:649-53.

53. Vajanapoom N, Shy CM, Neas LM and Loomis D. Associations of particulate matter and daily mortality in Bangkok, Thailand. *Southeast Asian J Trop Med Public Health.* 2002;33:389-99.

54. Roberts S. Interactions between particulate air pollution and temperature in air pollution mortality time series studies. *Environ Res.* 2004;96:328-37.

55. Bell ML, McDermott A, Zeger SL, Samet JM and Dominici F. Ozone and short-term mortality in 95 US urban communities, 1987-2000. *Jama.* 2004;292:2372-8.

56. Pope CA, 3rd, Hansen ML, Long RW, Nielsen KR, Eatough NL, Wilson WE and Eatough DJ. Ambient particulate air pollution, heart rate variability, and blood markers of inflammation in a panel of elderly subjects. *Environ Health Perspect.* 2004;112:339-45.

57. Peng RD, Dominici F, Pastor-Barriuso R, Zeger SL and Samet JM. Seasonal analyses of air pollution and mortality in 100 US cities. *Am J Epidemiol.* 2005;161:585-94.

58. Baccini M, Biggeri A, Accetta G, Lagazio C, Lerxtundi A and Schwartz J. Comparison of alternative modelling techniques in estimating short-term effect of air pollution with application to the Italian meta-analysis data (MISA Study). *Epidemiol Prev.* 2006;30:279-88.

59. Ren C, Williams GM, Mengersen K, Morawska L and Tong S. Does temperature modify short-term effects of ozone on total mortality in 60 large eastern US communities? An assessment using the NMMAPS data. *Environ Int.* 2008;34:451-8.

60. Assessment CoR. . *Science and Decisions: Advancing Risk Assessment.* . Washington, D.C.: National Academies Press; 2009.

---

[1] Moher D, Jones A, Lepage L, for the CONSORT Group for the C. Use of the CONSORT Statement and Quality of Reports of Randomized Trials. JAMA. 2001 Apr 18;285(15):1992.

[2] Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. PLoS Biol. 2010 Jun 29;8(6):e1000412.

[3] Dwan K, Altman D, Clarke M, Gamble C, Higgins J. Observational Studies: Getting Clear about Transparency. PLoS Med. 2014 Aug 26;11(8):e1001711.