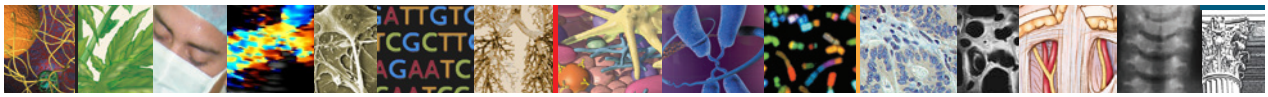

The NEW ENGLAND JOURNAL of MEDICINE

Perspective

"Transparency" as Mask? The EPA's Proposed Rule on Scientific Data

Joel Schwartz, Ph.D.

The Environmental Protection Agency (EPA) recently proposed excluding from consideration in setting environmental standards any studies whose raw, individual-level data are not publicly available. This proposal was preceded by the wholesale exclusion from the EPA's scientific advisory boards of academic scientists who receive research grants from the agency — and their replacement by industry-funded scientists. It is hard to interpret these actions as anything other than an attack on the use of hard scientific evidence to set environmental standards.

Open science has growing support, and justly so. However, studies conducted at academic institutions and involving humans, which are regulated by the Health Insurance Portability and Accountability Act (HIPAA) and institutional review boards (IRBs), must maintain a basic regard for privacy. Great progress in understanding pollution's effects has been made by adding exposure information to large cohort studies that were established to explore cardiovascular disease or cancer. Such studies have been used, for example, to analyze concentrations of metals in blood, urine, or toenails and to attribute air pollution exposure to people according to their residential address. Precisely because these studies include measurements of many potential confounding factors, it is difficult to make the data public without also making participants identifiable. Although some progress has been made in deidentifying some types of data, studies of environmental exposures present more serious issues, because often exposure levels are attributed on the basis of geocodes, and neighborhood covariates are based on public geocoded data. This practice makes it much easier to identify participants. For example, after Hurricane Katrina, a local newspaper published a map of the locations of deaths. It showed no roads, and the only geographic data included were neighborhoods. Yet researchers were able to correctly identify the residential address for most of the people who died.[1]

A cohort study of pollution rarely includes individual geocodes as covariates, but it typically controls for 15 to 20 potential confounders, usually including census-based measures of socioeconomic status (SES) and other geocoded information. If those covariates were all dichotomous, there would be more than 32,000 unique combinations. If some variables are based on publicly available geocoded data, such as census-tract measures of race, SES, population density, housing value, local air

pollution levels, and county-level data from the Behavioral Risk Factor Surveillance survey of the Centers for Disease Control and Prevention, it may be possible to identify each participant's census tract. With continuous confounders, the situation is worse. Identifiability is thus a major concern: if you know someone's age, race, sex, and other individual covariates, adding the census tract may make the participant unique, particularly if the outcome being studied is death, given that death certificates are obtainable.

This problem is well recognized: the National Academy of Sciences has reported on an "experiment to discover whether confidentiality could be preserved while opening . . . data for public review," which demonstrated that even after all participant features not required to allow other scientists to replicate a study's basic findings were deleted from study questionnaires, investigators could identify the participants.[2]

Recently, a study examined the identifiability of records from an environmental health study in Northern California. Using data considered under HIPAA to be sufficiently deidentified to be made public, they were able to correctly identify more than 25% of the participants.[3] Previously, the lead author showed that people from a supposedly anonymized hospital-admissions database could be identified on the basis of news stories. Since many obituaries are printed every day and death certificates are publicly available, the identifiability problem is vast.

In the Harvard Six Cities Study, conducted in the 1980s and 1990s, participants were recruited from one neighborhood in each city, including Watertown, Massachusetts (population, 35,000). The average number of deaths per year in Watertown is 208 — less than 1 death per day. Obviously, knowledge of the date of death would uniquely identify most participants. But even if the data made public included only the year of death, age, race, sex, and cause of death, most people could be identified from those facts.

The Canadian Community Health Survey followed 300,000 people and examined the association of exposure to fine particulate matter (particles with a mass median aerodynamic diameter of less than 2.5 μm [$PM_{2.5}$]) with mortality.[4] Because of privacy laws, the data were not given to the investigators, and analysis was performed on the computers at Statistics Canada. Yet this study is critical for the EPA to consider as it reviews the adequacy of its 12 μg-per-cubic-meter $PM_{2.5}$ standard, because essentially all the participants lived in locations with $PM_{2.5}$ levels below that standard.

The EPA's proposed rule on evidence for policymaking will exclude European and Canadian studies involving human participants from being considered by the EPA in regulating environmental pollutants. The new General Data Protection Regulation (GDPR) in the European Union (EU) defines private data as including information on a person's medical, physical, physiological, genetic, mental, economic, cultural, or social identity. Under the GDPR, such data must be controlled by a data controller who must demonstrate that any use of the data has been consented to by the individuals involved — which obviously precludes making data publicly available.

EPA leaders have argued that data can be sufficiently deidentified to be made public while still permitting reanalysis. But the number of variables included in original analyses that would have to be omitted or condensed into crude categories is so large that any reanalysis would be unable to reproduce the original results. More plausible is the EPA's argument that protected data centers could house the data and allow people to analyze them. But if the Canadian government would not allow the initial investigators to have the data mentioned above, then it's unlikely that it would agree to convey those data to an EPA computer, even with restricted access. Similar barriers probably apply to most of the cohort studies the EPA relies on: IRB and EU privacy rules are unlikely to allow transfer of data to EPA or other U.S. government computer centers.

Moreover, the "gold standard" of science is not reanalysis, but replication. In the case of $PM_{2.5}$-mortality studies, a recent meta-analysis found 53 cohorts, indicating that the results have been replicated many times by many groups in many countries.[5] Of what value, then, is a reanalysis of a minimal subset of covariates from any given study — particularly if it can't control for important covariates?

It is difficult to believe that EPA leaders do not know that few human cohort studies 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 data from consideration, resulting in a weakening of standards that are no longer supported by "sufficient scientific evidence." This approach was

outlined in a 1996 e-mail message, revealed in tobacco litigation, from a law firm to R.J. Reynolds. Addressing possible regulation of environmental tobacco smoke (ETS), it stated, "Because there is virtually no chance of [e]ffecting change on this issue if the focus is ETS, our approach is one of addressing process as opposed to scientific substance, and global applicability to industry rather than focusing on any single industrial sector." It highlighted ozone and $PM_{2.5}$ regulations as also ripe for this approach. Subsequently, polluting industries hired actors to stage a demonstration demanding that the Six Cities Study data be made public. Combined with the recent removal of impartial scientists from the EPA's scientific review boards, the current proposal appears to be a multipronged attack on the use of scientific data to set regulatory standards.

Rules suggesting that individual personal information might be made public can also endanger wider research on cancer, heart disease, and other conditions. People may be much less likely to agree to participate in long-term epidemiologic studies if they hear that they may be identified or their data made public. The EPA has not made a decision yet on this proposal and, I believe, should be encouraged to make one that preserves scientific input into its rulemaking.

Disclosure forms provided by the author are available at NEJM.org.

From the Departments of Environmental Health and Epidemiology, Harvard T.H. Chan School of Public Health, Boston.

This article was published on August 29, 2018, at NEJM.org.

1. Curtis AJ, Mills JW, Leitner M. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. Int J Health Geogr 2006;5:44
2. National Research Council. Access to research data in the 21st century: an ongoing dialogue among interested parties: report of a workshop. Washington, DC: National Academy Press, 2002.
3. Sweeney L, Yoo JS, Perovich L, Boronow KE, Brown P, Green Brody J. 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).
4. Pinault L, Tjepkema M, Crouse DL, et al. Risk estimates of mortality attributed to low concentrations of ambient fine particulate matter in the Canadian community health survey cohort. Environ Health 2016;15:18.
5. Vodonos A, Awad YA, Schwartz J. The concentration-response between long-term $PM_{2.5}$ exposure and mortality: a meta-regression approach. Environ Res 2018;166:677-89.

DOI: 10.1056/NEJMp1807751