MedicalXpress

Participants in environmental health studies vulnerable to re-identification

January 14 2020

Before sharing human research data, scientists routinely strip it of personal information such as name, address, and birthdate in order to protect the privacy of their study participants. However, reporting in the journal *Environmental Health Perspectives*, researchers at Silent Spring Institute and their colleagues show that for environmental health studies, that might not be enough—even anonymized data can sometimes be traced back to individuals.

The new study highlights the need for greater protections for participants in human research studies. It also has implications for a proposed federal rule by the U.S. Environmental Protection Agency (EPA) that would require scientists to make their data public in order for their research to be used as a basis for environmental regulations.

"Researchers promise to protect the privacy of their study participants—a routine practice in nearly all scientific studies involving people," says lead author Katherine Boronow, a staff scientist at Silent Spring. "Our research shows that making data publicly available from environmental health studies, even after obvious identifiers are removed, could violate these pledges."

In a previous study, Silent Spring researchers conducted an experiment in which they shared anonymized data from the Institute's Household Exposure Study in California with a team of Harvard researchers skilled in re-identification techniques. By linking housing and demographic data from the study to publicly-available data such as tax assessor records,

and using other information described in the study such as the location of the housing developments and the levels of indoor air pollutants measured, the team successfully re-identified 25 percent of participants from one housing development by name.

Now, in this latest investigation, the researchers show that vulnerability to re-identification is a common aspect of environmental health data. They reviewed a dozen environmental health studies and identified five different types of data (location, medical, genetic, occupation, and housing) that overlap with outside databases and could contribute to the risk of re-identification.

The researchers found that all 12 studies included at least two out of the five data types, and three studies included all five. "Having multiple data types provides more opportunities for someone to match research data against existing commercial or public databases," says Boronow.

Measurements of pollutants in people's bodies or in their homes are also a characteristic data type of many environmental health studies. Currently, however, these measurements alone are less vulnerable to data linkage because there are few databases that include chemical measurements that could be used for matching.

To explore a different way that chemical exposure data might be used in re-identification, the team conducted a cluster analysis using data from Silent Spring's Household Exposure Study in California and in Massachusetts and from the Centers for Disease Control's Green Housing Study in Boston and Cincinnati. They fed the raw chemical measurements to an algorithm that sorted the data within each study into two groups. The groups created by the algorithm corresponded to geographic location with 80 to 98 percent accuracy.

If the data cluster into groups by location, says Boronow, then each

group can be matched to data narrowed to that location, making it more likely for a re-identification attack to produce correct matches. This shows how someone could use chemical data to infer a characteristic of people in a study even if that characteristic is excluded when the study data are shared.

Data sharing has many benefits. By pooling data, researchers can create larger, more diverse datasets that could lead to advances in knowledge. It can also give researchers access to data that are difficult or expensive to obtain, such as data from biological or environmental samples collected after an environmental disaster. However, as the new study shows, it also has its risks.

Dr. Julia Brody, executive director at Silent Spring and a co-author of the study, says the implications of privacy risks are not trivial. Loss of privacy could result in stigma for individuals and communities. It could affect property values, insurance, or a person's chances of employment. It could also damage trust in research.

In 2018, EPA released a proposed rule called "Strengthening Transparency in Regulatory Science," that would require researchers to disclose their raw data as a precondition for the agency using a study to support regulatory decisions. Because the requirement could jeopardize confidential information about study participants, it could disqualify critical environmental health studies that form the basis of existing regulations, such as current limits on air pollutants. EPA is expected to release a revised version of the proposed rule early this year.

"Thousands of Americans have contributed personal data to scientific research with the goal of improving health for all," says Brody. "We must not take advantage of their generosity with rules that threaten their privacy and discourage future participation in research."

With growing pressure on scientists to share their data, and with more consumer data available online, Brody says it is important to fully characterize the risks of data sharing and identify solutions. Results from their research, she says, could help scientists develop informed consent documents that are more forthcoming about the risks and could help determine what types of data should be excluded from public sharing. It could also lay the groundwork for legal and policy protections for participants should they fall victim to re-identification.

More information: Katherine E. Boronow et al, Privacy Risks of Sharing Data from Environmental Health Studies, *Environmental Health Perspectives* (2020). DOI: 10.1289/EHP4817

Provided by Silent Spring Institute