LETTERS

Edited by **Jennifer Sills**

A new privacy debate

A STEADY STREAM of large-scale data breaches has focused attention on privacy and led to calls for anonymity, especially for collections of sensitive health data. Meanwhile, recent research has demonstrated—again—that true anonymization of an individual's data is virtually impossible ("Unique in the shopping mall: On the reidentifiability of credit card metadata," Y.-A. de Montjoye *et al.*, Reports, special section on The End of Privacy, 30 January, p. 536). Any policy focused on protecting patient privacy via anonymization will render crucial data useless for clinical and public health research, negate billions of dollars in data infrastructure and analysis investments, and cause real harm by slowing the pace of medical progress. Privacy policies geared to exceptions instead of the

norm, or that ignore the breadth and diversity of the many fields using identifiable data practice, will not be efficient.

We can protect individual privacy without sacrificing the potentially transformative insights that large collections of personally identifiable data provide. First, we must acknowledge that relying on anonymization algorithms to scrub our personal information from these data resources is not currently a viable solution. These methods can be circumvented by individuals both internal and external to the organization and anonymization process. Moreover, while failing to protect our identities, they also distort much of the information needed to make the data useful for fields such as medicine and public health.

Second, it is crucial that policies are sensitive to the different intents and practices of the many diverse fields using identifiable data. For example, the analysis of personally identifiable data in clinical and public health research is typically designed to statistically aggregate and compare large groups of people. Consider, for example, a statistical model built to predict a patient's likelihood of responding to cancer treatment. Such models are built by combining identifiable patient information from large numbers of individuals. The output of this research, however, is only valuable if it is generalizable beyond a single individual—de facto anonymization. These life-saving methods, therefore, can be used in practice without risking exposure of personally identifiable information. Sweeping policies that prevent these sorts of ethical and proper use cases risk derailing entire fields, such as public health and medicine.

Third, we must develop a stronger culture of individual participation along with greater transparency in activities that use individually identifiable data. In the health care field, for example, the Patient-Centered Outcomes Research Initiative (PCORI) (*1*, *2*), Agency for Healthcare and Research Quality (AHRQ) (*3*), and Institute of Medicine (IOM) (*4*) have been helping develop new policies regarding patient engagement and research dissemination, which are substantially changing the landscape regarding patient data. These funding initiatives should be expanded, and successful models should be widely shared.

Our debate, therefore, should not be focused on the efforts that protect privacy via anonymization. While research in that direction should continue, we must recognize that anonymization as a precondition to storing or collecting personal information is not a viable policy solution. Instead, we should focus our attention toward requiring safeguards on improper storage, distribution, or exploitation of personal data, and on developing a culture of trust and transparency surrounding the use of such data resources.

Anne-Marie Meyer*[1] *and David Gotz*[2]

[1]Gillings School of Global Public Health and Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [2]Department of Information Science and Carolina Health Informatics Program, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

*Corresponding author. E-mail: ameyer@unc.edu

REFERENCES

1. D. Hickam *et al.*, Eds., "The PCORI methodology report" (PCORI, 2013); www.pcori.org/assets/2013/11/PCORI-Methodology-Report.pdf.
2. PCORI, What We Mean By Engagement (www.pcori.org/content/what-we-mean-engagement).
3. Agency for Healthcare Research and Quality, "Findings and lessons from the enabling patient-centered care through health IT grant initiative" (Westat Under Contract No. HHSA 290200900023I, AHRQ Publication No. 13-0011-EF, Rockville, MD, 2013).
4. L. Olsen, R. S. Saunders, J. M. McGinnis, Eds., "Patients charting the course: Citizen engagement in the learning health system" (The Learning Health System Series, Institute of Medicine, 2011); www.iom.edu/Reports/2011/Patients-Charting-the-Course-Citizen-Engagement-in-the-Learning-Health-System-Workshop-Summary.aspx.

Assessing data intrusion threats

Y.-A. DE MONTJOYE *et al.*'s Report "Unique in the shopping mall: On the reidentifiability of credit card data" (special section on The End of Privacy, 30 January, p. 536) led to a widespread media sensation proclaiming that reidentification is easy with only a few pieces of credit card data (*1*–*3*). Although we agree with de Montjoye *et al.* that data disclosure practices must be responsibly balanced with data privacy and utility, we are concerned that the study's findings reflect unrealistic data intrusion threats. Making policy decisions based on the conclusions from this work would thus be hasty and could lead to the abandonment of modern data protection standards, with negative consequences to privacy, research, and society.

Some media confusion stems from the paper's use of the term "reidentify"; credit card metadata were not actually linked to any personal identities. Instead, it was assumed that an intruder could obtain data about identity, geography, time, and price to reidentify all targeted consumers. Yet this scenario requires some very strong assumptions about the attacker that are unlikely to be realized in practice. First, the study did not demonstrate the extent to which the necessary identifying information could be obtained reliably for any consumer. Second, the study neglects to acknowledge that when the data come from a fraction of the general population, unique purchase data in the sample will often not be unique in the larger population. Given that the undisclosed country's population was likely much larger than 1.1 million, the paper's data uniqueness measure is likely a substantial overestimate of risk. Third, the study's risk estimates are further inflated because they did not include cash or other banks' credit card purchases.

The research communicated in this paper is critical to moving forward privacy discussions about data sharing. However, we stress that claims about reidentification must be based on models that realistically and correctly account for the probability, as well as the possibility, of attacks.

Daniel Barth-Jones,*[1] ***Khaled El Emam,*[2]***

Xiangqian Li and Jiaqi Liu demonstrate a physical exam during the opening skit at Healthcare Day.

PHOTO: XIANGQIAN LI, ZHONGSHAN HOSPITAL, FUDAN UNIVERSITY

OUTSIDE THE TOWER

Acting to build trust

My 5-foot 9-inch, 160-pound frame hunches over in pain. "I am Macho Man," I declare to the crowd, "but because of my back pain, I can't even pick up a strawberry!" Suddenly, a white-coated angel appears beside me. I look up in awe. "Can you relieve my pain?" Thus begins the opening sketch at Healthcare Day in Shanghai, China.

Since 2012, I and others from Zhongshan Hospital at Fudan University have been holding Healthcare Day in Community, a quarterly event in which medical scientists from different departments provide health counseling and services for local residents.

As the skit continues, the white-coated angel explains the physiology behind back pain and shows me, Macho Man, how to prevent muscle strain. He models the correct way to lift a heavy load and demonstrates how a physical exam would diagnose the problem. Then he leaves the stage and walks through the audience, looking for other people in need.

During Healthcare Day, community members have the opportunity to ask scientists about genes, tissue engineering, cancer, and cutting-edge medical research. We advocate a healthy lifestyle and give relevant advice about exercise habits, weight control, and proper diet. This event builds trust between the community and medical scientists. We hope more medical scientists and doctors in China will join us in stepping out of the hospital and sharing health science with the public. ***Jiaqi Liu***

Department of Plastic Surgery, Zhongshan Hospital, Fudan University, 200032, Shanghai, China. E-mail: liujiaqi1213@yahoo.com

Jane Bambauer,[3] Ann Cavoukian,[4] Bradley Malin[5]

[1]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY 10032, USA. [2]College of Law, University of Arizona, Tucson, AZ 85721, USA. [3]Privacy and Big Data Institute, Ryerson University, Toronto, ON, M5B 2K3, Canada. [4]Children's Hospital of Eastern Ontario Research Institute and University of Ottawa, Ottowa, ON, K1H 8L1, Canada. [5]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37212, USA.

*Corresponding author. E-mail: db2431@Columbia.edu

REFERENCES

1. N. Singer, "With a few bits of data, researchers identify 'anonymous' people," *New York Times* (29 January 2015); http://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/.
2. R. Jacobson, "Your 'anonymous' credit card data is not so anonymous, study finds," *PBS NewsHour* (29 January 2015); www.pbs.org/newshour/rundown/anonymous-credit-card-data-anonymous-study-finds/.
3. D. Coldewey, "'Anonymous' credit card data can still give you away," *NBC News* (29 January 2015); www.nbcnews.com/tech/tech-news/anonymous-credit-card-data-can-still-give-you-away-n296446.

Response

BARTH-JONES *ET AL.* claim that our findings "reflect unrealistic data intrusion threats" We strongly disagree and argue that Barth-Jones *et al.*'s Letter is instead a superb illustration of why deidentification is not "a useful basis for policy" (*1*).

A simple and real example of our attack model is a bank sharing metadata for its 1.1 million customers in anonymized form with a third party for analysis. If the third party is able to obtain additional information—such as loyalty program data if the third party is a retailer—that data could be used to reidentify an individual and all the rest of his or her purchases.

Barth-Jones *et al.*'s Letter exemplifies the intrinsic issue with deidentification. One can always, as Barth-Jones *et al.* have, artificially lower the estimated likelihood of reidentification through the use of arbitrary and debatable assumptions.

First, Barth-Jones *et al.* have consistently considered an intrusion to be a breach of privacy only if "all targeted customers" are reidentified (*2*). This is an unrealistic definition of breach of privacy. Second, Barth-Jones *et al.* assume that it is "very unlikely" for an attacker to be able to collect geolocalized information about an individual. At best, this is a striking underestimation of the current availability of identified data. Possible sources would include manually collected clues about an individual we know (e.g., receipts or branded shopping bags) (*3*); having access or collecting from public profiles people's check-ins at shops or restaurants on Yelp, Foursquare, or Facebook (*4*); or having access to a retailer's database or to a database of geolocalized information such as the one collected by smartphone applications (*5*), WiFi companies, and virtually any carriers in the world. Third, Barth-Jones *et al.* assume that an attacker cannot know whether an individual is a client of a bank and is therefore in the data set. This is again an assumption that artificially lowers the estimated, and thus perceived, risks of reidentification without changing at all the actual risk for people in the release data set. Fourth, the fact that an individual might occasionally pay cash only means that an attacker would need a few more points.

Estimated probabilities of reidentification are not a useful basis for policy, and we stand by our comment that "the open sharing of raw [deidentified metadata] data sets is not the future" (*6*).

Yves-Alexandre de Montjoye* and Alex "Sandy" Pentland

Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*Corresponding author. E-mail: yvesalexandre@demontjoye.com

REFERENCES

1. President's Council of Advisors on Science and Technology, *Big Data and Privacy: A Technological Perspective* (PCAST, Washington, DC, 2014), pp. 38–39.
2. D. C. Barth-Jones, "Press and Reporting Considerations for Recent Re-Identification Demonstration Attacks: Part 2" (http://blogs.law.harvard.edu/billofhealth/2013/10/01/press-and-reporting-considerations-for-recent-re-identification-demonstration-attacks-part-2-re-identification-symposium/).
3. L. Sweeney, *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.* **10.05**, 557 (2002).
4. Wallaby, "Is anonymous financial data anonymous?" (www.walla.by/blog/110651700144/is-anonymous-financial-data-anonymous).
5. CNIL, "Mobilitics, season 2: Smartphones and their apps under the microscope" (www.cnil.fr/english/news-and-events/news/article/mobilitics-season-2-smartphones-and-their-apps-under-the-microscope/).
6. J. Bohannon, *Science* **347**, 468 (2015).

Assessing data intrusion threats--Response

Yves-Alexandre de Montjoye and Alex "Sandy" Pentland

Science **348** (6231), 195.
DOI: 10.1126/science.348.6231.195-a