Questioning the Limits of Genomic Privacy

To the Editor: Recently, Im et al.[1] presented a method that can infer an individual's participation in a study when regression coefficients from quantitative phenotypes are available. They demonstrated that in an era of increasing use of high-throughput technologies to integrate multiple-omics data sets, the "problem of identifiability" necessitates the creation of robust methods (e.g., an annual certification process) that facilitate broad dissemination of study results without compromising a participant's privacy. In this letter, we would like to qualify the conclusions of Im et al., and several other commentators,[2–5] by illustrating that (1) despite the perceived ease of reidentification, anonymity (and genomic privacy in general, which subsumes anonymity and identifiability as critical elements of informational control) remains a valid and vital concept and (2) technologies and models currently exist that facilitate dissemination of useful health data without compromising privacy. We think that the topic addressed by Im et al. is all the more critical given that the European Union (EU), the United States (US), and other jurisdictions are presently reforming their privacy, data, and human subjects research protection frameworks.

As policymakers, scientists, and the public grapple with the growing data deluge and concerns about privacy, a key issue will be to examine the legal definition of "personal data." The EU's newly proposed data protection regulation defines personal data as "any information relating to a data subject." A data subject is an "identified natural person" (i.e., a person whose identity data, such as name, address, or birth date, are known) or a "natural person who can be identified, directly or indirectly, by means reasonably likely to be used by... [a]... person."[6] A recent revision to the proposed regulation's definition of "personal data" adds that "[i]f identification requires a disproportionate amount of time, effort, or material resources, the natural living person shall not be considered identifiable."[7] In the US, according to the Health Insurance Portability and Accountability Act of 1996 (HIPAA), "individually identifiable health information" is information that identifies the individual or for which "there is a reasonable basis to believe it can be used to identify the individual."[8]

Neither the EU's proposed data protection regulation nor HIPAA provide definitions of "anonymous" or "anonymization," which have distinct technical meanings,[9] but nationally and internationally recognized definitions of "anonymous" exist, though they unfortunately continue to lack terminological and technical standardization.[3] For example, the EU's Article 29 Working Party defines anonymous data as "any information relating to a... person where the person cannot be identified... taking into account all means reasonably likely to be used."[10] To us, this is a clear recognition of the concept and utility of anonymous data. Yet, when it comes to biological data, like DNA parameters, many believe that anonymity simply no longer exists because the legal term "identifiable" seemingly now applies to everyone because every "anonymous" or "anonymized" person can sooner or later be identified by some technology and method.

This argument overlooks many critical points. First, a biospecimen in itself does not contain identity data. Even if it can be determined with a certain probability that a biospecimen originates from a specific individual by matching DNA data, such matching is different from assessing the identity of an individual.[11,12] Furthermore, the more uncertainty there is in determining data for reidentification, the more anonymous the data become; absent true data authenticity, reidentification risks are minimal.[13] Even when reidentification on the basis of deidentified or anonymized biomedical data would be possible because databases with voter registration data, hospital discharge data, and court proceedings are accessible, a survey showed that reidentification on the basis of properly "deidentified" (to say nothing of anonymous) data is extremely difficult to achieve in practice.[14,15] In sum, lending unreasonable credibility to remote risks of reidentification confuses multiple, justifiably separate legal definitions of "personal data," "data subject," "anonymous," and "anonymized" and leads to a burdensome "gross overexpansion of the [privacy] legal framework."[16] This in turn threatens the advancement of anonymity as a practical concept, curtails beneficial uses of data, and reduces the incentive to anonymize data or collect anonymous data.[17] In both science and in law, then, data anonymity vitally remains an ongoing concern. Remote exceptions cannot form the basis for a common rule. Data is not "personal" if "anonymous" or "anonymized."

Second, similar to our objections to those who treat all data as "personal," we think that there is a widespread failure to accept the rapid technological progress being made, particularly in genomics research and population biobanking, to simultaneously protect an individual's privacy interests and promote scientific and biomedical breakthroughs.[18–20] Current practices such as data access agreements already incorporate the annual certification process that Im et al. propose.[21] There are ample reasons to move past the stale dichotomy and false choice of privacy or data utility and to embrace the possibilities of emerging technologies, processes, and projects. Far from potentially harming participants and researchers, methods and emerging technologies that work within a regulatory framework or legislation demonstrate how anonymity may facilitate innumerable benefits.

Certainly privacy protection remains the most pressing concern within the interface of medical research and public participation. Indeed, there are areas that warrant greater focus by the scientific community, such as group-based privacy issues where, for example, "nontransparent allocation of individuals to groups based on known or inferred traits or some combination thereof can raise issues related to the ability to protect one's own interest and avoid discrimination."[22] We share the concern of Im et al. and others that as science and technology advance, the use of additional human characteristics such as data will pose challenges to privacy interests, which may need to be reconceptualized to remain relevant in 21st century science and medicine. Yet, concerns regarding the "problem of identifiability" as a veritable limit to genomic privacy must be tempered with nuance. It is only through recognition and acceptance of the ongoing practical utility of data anonymity, use of evidence to conclude that the risk of reidentification is remote, and adoption of successful emerging practices and technologies that we can achieve a "win-win" situation. Anonymous and useful data can be legally and ethically bridged while respecting the privacy interests of individual participants, along with the biomedical research interests of society as a whole.

Bartha M. Knoppers,[1,*] Edward S. Dove,[1] Jan-Eric Litton,[2] and J.J. Nietfeld[3]
[1]Centre of Genomics and Policy, Department of Human Genetics, Faculty of Medicine, McGill University, Montreal, QC H3A 0G1, Canada; [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A, 171 77 Stockholm, Sweden; [3]Department of Pathology, University Medical Center Utrecht, Heidelberglaan 100, 3584CX Utrecht, The Netherlands
*Correspondence: bartha.knoppers@mcgill.ca

References

1. Im, H.K., Gamazon, E.R., Nicolae, D.L., and Cox, N.J. (2012). On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am. J. Hum. Genet. *90*, 591–598.
2. Craig, D.W., Goor, R.M., Wang, Z., Paschall, J., Ostell, J., Feolo, M., Sherry, S.T., and Manolio, T.A. (2011). Assessing and managing risk when sharing aggregate genetic variant data. Nat. Rev. Genet. *12*, 730–736.
3. Schmidt, H., and Callier, S. (2012). How anonymous is 'anonymous'? Some suggestions towards a coherent universal coding system for genetic samples. J. Med. Ethics *38*, 304–309.
4. Ohm, P. (2010). Broken promises of privacy: responding to the surprising failure of anonymization. UCLA Law Rev. *57*, 1701–1777.
5. Sándor, J., and Bárd, P. (2011). The question of anonymity and privacy in biobanking. In Biobanks and Tissue Research: The Public, the Patient and the Regulation, C. Lenk, J. Sándor, and B. Gordijn, eds. (New York: Springer), pp. 213–230.
6. European Commission (2012). Proposal for a regulation of the European Parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data. EC's proposal, http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.
7. Council of the European Union (2012). Revised proposal for a regulation of the European Parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Council's revised proposal, http://www.statewatch.org/news/2012/jun/eu-council-revised-dp-position-11326-12.pdf.
8. Health Insurance Portability and Accountability Act of 1996 (HIPAA). U.S. 45 C.F.R. § 160.103.
9. Knoppers, B.M., and Saginur, M. (2005). The Babel of genetic data terminology. Nat. Biotechnol. *23*, 925–927.
10. Article 29 Working Party (2007). Opinion 4/2007 on the concept of personal data. Working party opinion, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf.
11. Nietfeld, J.J. (2007). What is anonymous? EMBO Rep. *8*, 518.
12. Malin, B., Loukides, G., Benitez, K., and Clayton, E.W. (2011). Identifiability in biobanks: models, measures, and mitigation strategies. Hum. Genet. *130*, 383–392.
13. Masiello, B., and Whitten, A. (2010). Engineering privacy in an age of information abundance. AAAI Spring Symposium Series 119–124, https://www.aaai.org/ocs/index.php/SSS/SSS10/paper/view/1188/1497.
14. El Emam, K., Jonker, E., Arbuckle, L., and Malin, B. (2011). A systematic review of re-identification attacks on health data. PLoS ONE *6*, e28071.
15. Malin, B., Karp, D., and Scheuermann, R.H. (2010). Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J. Investig. Med. *58*, 11–18.
16. Tene, O. (2011). The complexities of defining personal data: anonymization. Data Prot. Law Policy *8*, 6–7.
17. Tene, O., and Polonetsky, J. (2012). Privacy in the age of big data: a time for big decisions. Stanford Law Rev. Online *64*, 63–69.
18. Murtagh, M.J., Demir, I., Jenkings, K.N., Wallace, S.E., Murtagh, B., Boniol, M., Bota, M., Laflamme, P., Boffetta, P., Ferretti, V., and Burton, P.R. (2012). Securing the data economy: translating privacy and enacting security in the development of DataSHIELD. Public Health Genomics *15*, 243–253.
19. Ohno-Machado, L., Bafna, V., Boxwala, A.A., Chapman, B.E., Chapman, W.W., Chaudhuri, K., Day, M.E., Farcas, C., Heintzman, N.D., Jiang, X., et al; iDASH Team. (2012). iDASH: integrating data for analysis, anonymization, and sharing. J. Am. Med. Inform. Assoc. *19*, 196–201.
20. Nietfeld, J.J., Sugarman, J., and Litton, J.E. (2011). The Bio-PIN: a concept to improve biobanking. Nat. Rev. Cancer *11*, 303–308.
21. Joly, Y., Dove, E.S., Knoppers, B.M., Bobrow, M., and Chalmers, D. (2012). Data Sharing in the Post-Genomic World: The Experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). PLoS Comput. Biol. *8*, e1002549.
22. Heeney, C., Hawkins, N., de Vries, J., Boddington, P., and Kaye, J. (2011). Assessing the privacy risks of data sharing in genomics. Public Health Genomics *14*, 17–25.

http://dx.doi.org/10.1016/j.ajhg.2012.05.027.