The Royal New Zealand College of General Practitioners
Te Whare Tohu Rata o Aotearoa

Research using electronic health records: not all de-identified datasets are created equal

Vithya Yogarajan MSc Hons (Appl Math), MSc Hons (Comp Sci);[1]
Rajan Ragupathy BSc, BPharm, PhD[2,3]

[1]Department of Computer Science, The University of Waikato, Hamilton, New Zealand
[2]Pharmacy Services, Waikato District Health Board, Hamilton, New Zealand
[3]Corresponding author. Email: rajan.ragupathy@gmail.com

J PRIM HEALTH CARE
2019;11(1):14–15.
doi:10.1071/HC19010
Received 04 February 2019
Accepted 12 February 2019
Published 03 April 2019

We read the article *Research using electronic health records: balancing confidentiality and public good* by Wallis *et al.* with great interest. The authors note general practices need to trust de-identification processes when releasing patient records.[1] Patients have also expressed concerns about de-identification practices.[2] De-identification encompasses a wide range of practices, and there are no universally accepted standards.[2,3] We propose here a three-step scheme for judging de-identified health records: (1) the de-identification standards used (2) the performance of the de-identification system and (3) additional security measures taken to prevent re-identification. Such a scheme may be useful to ethics committees, researchers planning a project and health providers deciding whether to participate.

De-identification standards

The United States Health Insurance Portability and Accountability Act 1996 (HIPAA) provides arguably the most user-friendly definition of de-identified. Under HIPAA's Safe Harbor provision, 18 specific categories of protected health information (PHI) about patients and family members need to be removed from the records.[4] The New Zealand Health Information Privacy Code requirement that the information is in a form in which the individual is not identified is less specific, but arguably provides researchers greater flexibility.[3,5] However, the European Union's General Data Protection Regulation (GDPR) is arguably even more stringent than the HIPAA, and has extra-territorial reach. It requires that individuals are not identifiable rather than simply not identified (eg through cross-matching with other datasets or publically available information).[6,8]

Performance of the de-identification system

De-identification is a two-step process where PHIs are identified and replaced by appropriate surrogates. Recently, there have been significant advances in automating de-identification of health records using machine learning. Several systems have achieved the gold standard of 95% accuracy in identifying HIPAA Safe Harbor PHIs.[9] However, there are still challenges and concerns in automating the surrogate generation and replacement process. There are also concerns about the usability of records de-identified to this extent, and whether analysis of de-identified records will produce the same results as records that have not been de-identified.

Additional security measures

These include encryption, random noise generation and compartmentalisation of the datasets. Such measures protect de-identified data from being re-identified through cross-matching with other datasets.[7,8] A multi-layered protection model based on well-accepted patient safety practices may be useful.[10]

In conclusion, de-identification may more accurately be described as difficulty in identifying, and lies on a spectrum from very easy to near impossible. Being specific about where one's dataset lies allows researchers and health providers to make informed choices.

CSIRO Publishing

Competing interests

The authors declare no competing interest.

References

1. Wallis K, Eggleton K, Dovey S, et al. Research using electronic health records: balancing confidentiality and public good. J Prim Health Care. 2018;10(4):288–91. doi:10.1071/HC18040
2. OKeefe CM, Connolly CJ. Privacy and the use of health data for research. Med J Aust. 2010;193(9):537–41.
3. Yogarajan V, Mayo M, Pfahringer B. Privacy protection for health information research in New Zealand district health boards. N Z Med J. 2018;131:(1485):19–26.
4. United States Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. HHS.gov. [cited 2019 January 31]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
5. Office of the Privacy Commissioner. Comparison paper on health privacy laws. Privacy Commissioner. [cited 2019 January 31]. Available from: https://www.privacy.org.nz/news-and-publications/books-and-articles/comparison-paper-on-health-privacy-laws-2/
6. Rumbold JMM, Pierscionek B. The effect of the general data protection regulation on medical research. J Med Internet Res. 2017;19(2):e47. doi:10.2196/jmir.7108
7. Brasher E. Addressing the failure of anonymization: Guidance from the European Unions General Data Protection Regulation. Columbia Business Law Review Vol. 2018, Issue 3, 2018. [cited 2019 January 31]. Available from: https://cblr.columbia.edu/addressing-the-failure-of-anonymization-guidance-from-the-european-unions-general-data-protection-regulation/
8. Polonetsky J, Tene O, Finch K. Shades of gray: seeing the full spectrum of practical data de-identification. Santa Clara Law Rev. 2016;56(3):593–629.
9. Yogarajan V, Pfahringer B, Mayo M. Automatic end-to-end de-identification: is high accuracy the only metric? Computers and Society, Cornell University. arXiv:1901.10583 [cs.CY]. [cited 2019 January 31]. Available from: https://arxiv.org/pdf/1901.10583.pdf
10. Ragupathy R, Yogarajan V. Applying the Reason Model to enhance health record research in the age of big data. N Z Med J. 2018;131(1478):65–7.

Response

Thank you for putting forward this interesting suggestion. Having a score that rates the level of de-identification of health information could assist communication about de-identification and would potentially be of interest to researchers, patients, and practices. However, the development of such a scoring system is some time away. In the meantime, we need to continue to work to improve the reliability of current de-identification processes.

Katharine Wallis, MBChB, PhD, MBHL, Dip Obst, FRNZCGP
Department of General Practice & Primary Health Care
Bldg 730-380, 261 Morrin Rd, Auckland 1072
University of Auckland, New Zealand