The rise of big data in healthcare research

How can we address public concerns of privacy?

Richard Yu, Lilian Robinson, Salonee Patel

ABSTRACT

Big data, an emerging technological field which encompasses the rapid collection and analysis of massive datasets in order to generate and characterize patterns, is currently utilized in many industries ranging from banking to marketing. Its role in the healthcare field is currently being explored, and it has the potential to greatly improve healthcare and disease surveillance through pattern analysis of health data. Concerns held by the general public primarily focus on potential breaches of privacy and confidentiality of patient medical health records in the context of research. These concerns relate to the innate characteristics of big data, such as large size and fast data acquisition speed, which increases the risk of breaching confidentiality due to increasing reliance on digital cloud computing and increasing access to safe haven information. Therefore, it is important for physicians to be mindful of privacy concerns as big data becomes more prominent, as doing so is a key factor in building public trust in the use. Understanding strategies and limitations of current practice standards will allow physicians to build on existing guidelines to facilitate the rise of big data. This entails prioritizing privacy when handling big data through anonymization, creating safe havens, and promoting dynamic informed consent practice standards.

INTRODUCTION TO BIG DATA - THE 4 V'S

Big data describes the generation, capture, and analysis of large, complex data sets to provide useful trends and patterns. It is characterized by the 4 V's: volume, velocity, variety, and veracity.^{1,2} The first three V's relate to the rapid rise of digital and computational technology, leading to the production of massive datasets from multiple different sources that are collected in real time.^{1,2} Veracity refers to the trustworthiness of the data, meaning data collection and analysis should be structured to ensure completeness of data and reliability.^{1,2} Collectively, the 4 V's allow for predictive analyses used by many industries to guide their decisions in improving their services or products. For example, within the marketing industry, major corporations such as Walmart can generate up to 2 petabytes of data per day. The data varies from transactions to online activity and is analyzed to determine consumer behaviours.³

The application of big data within the healthcare system has also been of increasing importance and interest, due to its potential for improving patient management. For example, the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), a database of electronic medical records (EMR) that utilizes big data analytics,

have demonstrated the importance of chronic disease surveillance which lead to improvements in patient management.^{4,5} It is estimated that there is a total of 150 exabytes of data within the healthcare sector, with a growth of 2 exabytes every year.⁶ In the modern healthcare setting, there is a vast amount of data that can be collected for a single person, such as patient history records, medical prescriptions, investigations, and physical exams. By integrating all data at a public health level, physicians and researchers can easily track and study population health, and conduct disease surveillance in more detail than ever before.^{1,7}

With the rise of big data, the potential to find new insights into disease processes through pattern mining can be of significant value for the development of new management guidelines, treatment options, and preventative protocols which ultimately reduce healthcare costs.⁸⁻¹¹ For example, recent studies analyzing the CPSSCN database allowed for the generation of algorithms to identify patients at risk for dyslipidemia,¹² and development of quality management strategies to reduce healthcare costs.¹³ However, to facilitate these developments, confidential information must be shared amongst healthcare professionals.¹¹ Therefore, valid concerns exist regarding the collection and sharing of such large datasets, particularly for research. To address these concerns, we discuss public concerns relating to big data, as well as the role physicians have in managing this emerging technology and minimizing harm for patients in the context of research.

PUBLIC CONCERNS OF BIG DATA

Confidentiality is a core tenant of the patient-physician relationship, described long before the existence of EMRs.¹⁴ In Canada, personal information is protected under the Canada Privacy Statutes. While these statutes encompass big data, no binding guidelines exist that apply to its use in healthcare and research.¹⁵ Despite the limited guidance provided by these statutes, physicians must exercise caution in protecting patient health information in conjunction with the rise of big data. Such caution is integral to the maintenance of public trust in both electronic records and the medical profession as a whole. In the modern context of electronic healthcare data, confidentiality is of particular importance to the general public.¹⁶

Currently, there is conditional acceptance for the collection and sharing of electronic healthcare data in research among the general public.¹⁶ Despite general acceptance, trust in electronic data sharing in the healthcare sector is nuanced. For instance, patients indicate increased trust in the public sector relative to the private sector due to higher levels of accountability and data security.

¹⁷ Any association of electronic data with research initiatives

designed for generating profit is also deemed less trustworthy by patients.¹⁷ Importantly, when physicians adhere to the standards of professionalism and accountability, patients are more comfortable with the sharing of their data for the benefit of public interest.¹⁷

Information privacy protection influences the public opinion of big data; however, preservation of said privacy is not straightforward. For example, collection of genomic data at a population level can generate a vast amount of data, up to 40 petabytes per year. It is impractical to store this amount of data using raw hardware.¹¹ Instead, more and more data are being deposited onto cloud storage. While this practice facilitates ease of file transfer and collaboration, many strict regulations must be put in place by institutions in order to preserve patient privacy and protect personal health information.¹¹

OUR ROLES AS PHYSICIANS

It is important to note the dichotomy of physician and researchers in how they interact with patient data. Physicians, as data custodians, are tasked with maintaining privacy and security of the data to maintain public trust.^{17,18} Researchers, such as epidemiologists, work to anonymize data and perform the analyses.¹⁹ With the rise of big data, clinicians now have access to vast amounts of personal and population health information, placing increased emphasis on ethical decision making in routine healthcare and research.^{18,20} When physicians obtain confidential information, they have an ethical duty to maintain confidentiality and thereby demonstrate respect and cultivate trust from their patients. Thus, any breach in confidentiality may significantly hinder the patient-physician relationship.^{18,20}

Ethical duties must be balanced with a physician's professional duties in the context of clinical research.²⁰ One strategy physicians can implement to maintain privacy is to anonymize patient data by deidentifying their personal information in the context of a research study. Anonymization is a simple process achieved by removing any identifying information such as names, geographic data, and dates.^{20,21} Despite efforts to implement anonymization, big data creates increased risk of data leakage and makes anonymization challenging due to the overwhelming volume and points of data from multiple sources. Additionally, as computational power increases with advances in technology, the ability to keep data anonymized becomes increasingly challenging.²² Accidental or unauthorized disclosure could lead to negative consequences for patients involved. Therefore, it is important for the physician to minimize risk of accidental disclosure in the context of research and to disclose potential risks to patients.

Another important consideration for protecting private healthcare data is to safeguard the information appropriately. Thus, the creation of safe havens, trusted databases like CPSCCN, that only verified professional institutions have access to, are of utmost importance as they limit the loss of information.^{20,21} To access safe haven data, researchers and physicians must submit a proposed research plan and any risk of data breach must be disclosed to patients. However, different countries and institutions

have discrepant guidelines and vary in their capacity to effectively preserve privacy.^{11,22} There is also current limitations in connectivity to safe haven networks as not all primary care physicians can connect to hospital data.²² As more people gain access to information online, even within the context of safe havens, the risk of breaking confidentiality is increased.^{20,21} Since big data in healthcare is relatively new, it is imperative to calibrate existing guidelines in the context of big data. For example, physicians participating in research should require legal authority to collect and manage big data, and ethics board internal reviewing processes should be updated to adequately assess the use of big data within proposed projects.²⁰⁻²²

To further safeguard patient data, researchers must assess any potential privacy risks and disclose these upfront to the patient. To maintain best practice standards, it is recommended that physicians allow for dynamic informed consent from the patient, although it is important to note that not all research studies may require this level of consent.^{23,24} Dynamic consent is a continuous, documented, ongoing dialogue between the patient and physician about the data of interest being used in a study. Furthermore, the use of data disclosure agreements should be considered to minimize risk of disclosure by keeping the physician accountable.^{23,24} In the rare event of reidentification, accidental or unauthorized, patients must be counseled at the start of the study such that they understand all aspects of the protocols and are warned of the potential risks of having their information disclosed to a third party during the process of the study.^{23,24}

CONCLUSION

Big data is an emerging technology within healthcare which is associated with benefits for improving patient outcomes and disease surveillance as evidenced by CPSCCN in monitoring chronic diseases. Public concern pertaining to the use of big data stems mainly from issues of privacy and security of personal information. In general, the public demonstrates support for the collection and sharing of information between researchers and physicians, often citing trust towards the physicians as crucial. Therefore, physicians should emphasize the importance of trust by ensuring confidentiality and privacy are maintained throughout the research project. This can be achieved by way of data deidentification, safe havens for trusted access, and developing and promoting dynamic informed consent practice standards. Even with the advancement of technology and as people become more comfortable with EMRs and big data analytics, the emphasis on privacy and security remains a legitimate public concern, and physicians play an important role in maintaining that trust throughout research and clinical practice.

REFERENCES

1. The Canadian Medical Protective Association. The impact of big data on Healthcare and medical practice [Internet]; The Canadian Medical Protective Association; 2014 [cited 2018 Mar 12]. Available from: https://www.cmpa-acpm.ca/static-assets/pdf/research-and-policy/public-policy/com_14_big_data_design-e.pdf.

2. Kitchen R, McArdle, G. What makes big data, big data? exploring the ontological characteristics of 26 datasets. *Big Data Soc.* 2016 Feb;3:1-10. <https://doi.org/10.1177/2053951716631130>
3. Erevelles S, Fukawa N, Swayne L. Big data consumer analytics and the transformation of marketing. *J Bus Res.* 2016 Feb;69(2):897-904. <https://doi.org/10.1016/j.jbusres.2015.07.001>
4. Williamson T, Green ME, Birtwhistle R, *et al.* Validating the 8 CPCSSN Case Definitions for Chronic Disease Surveillance in a Primary Care Database of Electronic Health Records. *Annals of Family Medicine.* 2014 Jul;12(4):367-72. <https://doi.org/10.1370/afm.1644>
5. Garies S, Birtwhistle R, Drummond N, *et al.* Data Resource Profile: National electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), *Int J Epidemiol.* 2017 Aug;46(4):1091-1092f. <https://doi.org/10.1093/ije/dyw248>
6. Bellazzi R. Big Data and Biomedical Informatics: A Challenging Opportunity. *Yearb Med Inform.* 2014 May;9(1):8-13. <https://doi.org/10.15265/IY-2014-0024>
7. Canadian Health Infoway. The emerging benefits of electronic medical record use in community-based care [Internet]. Canadian Health Infoway; 2013. [cited 2018 Mar 12]. Available from: <https://www.infoway-inforoute.ca/en/component/edocman/1225-the-emerging-benefits-of-electronic-medical-record-use-in-community-based-care-executive-summary/view-document?Itemid=0>.
8. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol.* 2016 Jun;13(6):350-9. <https://doi.org/10.1038/nrcardio.2016.42>
9. Fillmore C, Braye C, Kawamoto, K. Systematic review of clinical decision support interventions with potential for inpatient cost reduction. *BMC Med Inform Decis Mak.* 2013 Dec;13:135. <https://doi.org/10.1186/1472-6947-13-135>
10. Jaspers M, Smeulders M, Vermeulen H, *et al.* Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *J Am Med Inform Assoc.* 2011 May;18(3):327-34. <https://doi.org/10.1136/amiainjnl-2011-000094>
11. Eisenstein M. Big data: The power of petabytes. *Nature.* 2015 Nov;527(7576):S2-4. <https://doi.org/10.1038/527S2a>
12. Aref-Eshghi E, Oake J, Godwin M, *et al.* Identification of Dyslipidemic Patients Attending Primary Care Clinics Using Electronic Medical Record (EMR) Data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) Database. *J Med Syst.* 2017 Mar;41(3):45. <https://doi.org/10.1007/s10916-017-0694-7>
13. Greiver M, Drummond N, Birtwhistle R, *et al.* Using EMRs to fuel quality improvement. *Can Fam Physician.* 2015 Jan;61(1): 92, e68-9.
14. Alvarez RC. The promise of e-Health – a Canadian perspective. *eHealth Int.* 2002 Sep;1(1):4. <https://doi.org/10.1186/1476-3591-1-4>
15. Osler, Hoskin & Harcourt LLP. International Comparative Legal Guide to: Data Protection 2016 [Internet]. London: Global Legal Group Ltd; 2016. p50-59. [cited 2018 Mar 12]. Available from: <https://www.osler.com/osler/media/Osler/reports/privacy-data/Data-Protection-Legislation-in-Canada.pdf>.
16. Aitken M, de St. Jorre J, Pagliari C, *et al.* Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Med Ethics.* 2016 Nov;17(1):73. <https://doi.org/10.1186/s12910-016-0153-x>
17. Stevenson F, Lloyd N, Harrington L, *et al.* Use of electronic patient records for research: views of patients and staff in general practice. *Fam Pract.* 2013 Apr;30(2):227-32. <https://doi.org/10.1093/fampra/cms069>
18. van Staa TP, Goldacre B, Buchan I, *et al.* Big health data: the need to earn public trust. *BMJ.* 2016 Jul;354:i3636. <https://doi.org/10.1136/bmj.i3636>
19. Capron AM. Protection of research subjects: do special rules apply in epidemiology? *J Clin Epidemiol.* 1991;44 Suppl 1:81s-89S. [https://doi.org/10.1016/0895-4356\(91\)90180-H](https://doi.org/10.1016/0895-4356(91)90180-H)
20. Panel on research Ethics. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans [Internet]. Government of Canada; 2014. [cited 2018 Mar 12]. Available from: http://www.pre.ethics.gc.ca/pdf/eng/tcps2-2014/TCPS_2_FINAL_Web.pdf.
21. Information and Privacy Commissioner of Ontario. Embedding Privacy into the Design of EHRs to Enable Multiple Functionalities – win/win. Information and Privacy Commissioner of Ontario; 2012. [cited 2018 Mar 12]. Available from: https://www.ipc.on.ca/wp-content/uploads/Resources/pbd-ehr-e_1.pdf.
22. Kulnych J, Greely HT. Clinical genomics, big data, and electronic medical records: reconciling patient rights with research when privacy and science collide. *J law Biosci.* 2017 Jan;4(1):94-132. <https://doi.org/10.1093/jlb/lsw061>
23. Riordan F, Papoutsis C, Reed J, *et al.* Patient and public attitudes towards informed consent models and levels of awareness of electronic health records in the UK. *Int J Med Inform.* 2015 Apr;84(4):237-47. <https://doi.org/10.1016/j.ijmedinf.2015.01.008>
24. Schadt EE. The changing privacy landscape in the era of big data. *Mol Syst Biol.* 2012;8:612. <https://doi.org/10.1038/msb.2012.47>