The Potential Use of Big Data in Cardiovascular Research

Anna Chu, MHSc[1,2]; Anam M. Khan, MPH[1]; Jack V. Tu, MD, PhD[1,3,4]

[1]Institute for Clinical Evaluative Sciences, Toronto, ON
[2]Department of Physical Therapy, University of Toronto
[3]Department of Medicine, Institute of Health Policy, Management and Evaluation, Institute of Medical Science, University of Toronto
[4]Schulich Heart Centre, Sunnybrook Health Sciences Centre

Over the last decade, interest in the field of big data has been increasing exponentially in all industries with new technological and methodological capabilities to collect, process, store, analyze and interpret the vast amount of data produced in almost every aspect of people's lives. The data that is generated is frequently used by companies to improve business processes, gain competitive advantages and inform risk management decisions. However, in health care, the use of big data is less developed compared to the commercial and financial industries. Moreover, Canada has yet to fully capitalize on its availability, lagging behind the United States and countries in Europe where big data is being used to improve evidence-based medicine, quality of medical care and health care spending efficiency. Canada has some of the world's most comprehensive population-based health administrative databases and clinical registries from data routinely collected in its universal health care system. Together with the growing use of electronic medical records, the ability to link together different data sources with varying information, using unique encrypted identifiers affords researchers the potential ability to develop very large study cohorts while saving on the costs associated with primary data collection. The wealth of information available on large study populations provides significant advantages for research, such as a greater ability to study sub-populations and rare conditions, as well as subsequent implementation of results.

An early Canadian example of the potential of 'big data' is the CArdiovascular HEalth in Ambulatory care Research Team (CANHEART) initiative, aimed at measuring and improving the cardiovascular health of Canadians and the delivery of quality ambulatory cardiovascular care in Ontario, Canada (URL in Appendix).[1] In this initiative, we have linked 17 population-based health administrative, vital statistic, survey, clinical, laboratory, drug, and electronic medical record databases available at the Institute for Clinical Evaluative Sciences (ICES) using encrypted personal identifiers to create a database of 9.8 million individuals, or almost the entire adult population of Ontario.[1] Information available in the CANHEART database includes socio-demographics, behavioural and traditional cardiac risk factors, health care utilization, comorbidities, drug use and clinical measures and outcomes. The creation of big data cohorts such as CANHEART gives us unique opportunities to enhance clinical research and evidence-based medicine to identify knowledge gaps, better inform clinical decisions, and personalize care.

Using the CANHEART database, we found striking 4-fold variations in the incidence of cardiovascular events among 800,000+ immigrants to Canada born in 201 countries around the world, from eight major ethnic groups. This is partly attributable to differences in the prevalence of traditional cardiac risk factors.[2] More recently, we reported on the cardiovascular risk of 5.5 million adults aged 40-79 years old across Ontario's 14 health service regions. We found a 2-fold difference in cardiovascular risk between the highest and lowest regions. Additionally, 33% of the variation between regions was due to differences in the prevalence of traditional risk factors, 25% was due to socio-demographics including ethnicity, and a further 16% was due to differences in the use of preventive cardiovascular care.[3] Using these findings, our team has developed an interactive online tool which allows users to visually compare data on cardiovascular indicators across these health service regions (URL in Appendix). Examples such as these demonstrate that – despite significant progress in cardiovascular health and health care – disparities in cardiovascular health exist and gaps in knowledge and health care remain. However, they also highlight some of the countless possible uses of big data in the future.

Beyond linkages of only health-related data, combining this data with non-traditional data (e.g., from activity trackers, smartphone applications, genetic studies and environmental data sources) provides further opportunities to improve cardiovascular outcomes. One such opportunity lies in the realm of personalized medicine. By leveraging and linking these information-rich data platforms from large populations, we may be able to develop enhanced cardiovascular risk prediction models that account for emerging risk factors and more accurately stratify patients at risk when compared to current cardiovascular risk prediction algorithms which are based on a limited number of patient characteristics and derived from smaller cohorts. The potential value of machine learning in

Corresponding Author:
Jack V. Tu, MD, PhD
Institute for Clinical Evaluative Sciences (ICES)
G106-2075 Bayview Ave, Toronto, ON M4N 3M5
E-mail: tu@ices.on.ca
Tel: 416-480-4700   Fax: 416-480-6048

the analysis of large datasets for, amongst other things, outcome prediction has also gained traction recently and is currently being explored.

Big data also has the potential to alter the way we have traditionally conducted randomized controlled trials (RCT). Considered the 'gold standard' of study designs, RCT findings have resulted in many advances in the treatment of cardiovascular disease, including the use of lipid-lowering statin therapy and decisions about percutaneous coronary intervention (PCI) versus bypass surgery.[4,5] However, use of big data in 'randomized registry trials' (RRT) is currently receiving attention as a more efficient and cost-effective method of conducting clinical trials. Rather than recruiting eligible patients (e.g. using recruitment ads), collecting self-reported information and physical and/or biological measures from them at study visits and then following them for a set time period for outcomes, as is often the case in traditional RCTs, patients in a RRT are recruited from existing registries already containing much of the data required for study. By facilitating patient recruitment, measurement of baseline patient characteristics, and uptake of trial interventions by trial participants for whom data is already collected in an administrative database, clinical registry or electronic medical record, use of big data for RRTs could significantly reduce the costs of conducting clinical trials, allow for longer follow-up times and increase the generalizability of the resulting sample.[6,7] One example is the Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia (TASTE) trial examining whether intracoronary thrombus aspiration improves flow and myocardial perfusion.[8] A diverse and broad cohort of patients in the Swedish angiography and angioplasty registry (SCAAR) platform were randomized to receive either conventional PCI or thrombus aspiration followed by PCI, with outcomes ascertained using SCAAR and other national registries. While already being used in some countries around the world, RRTs remain largely unchartered territory in Canada.

Use of big data in health care comes with enormous opportunities; however, it also presents with many challenges. Valuable personal-level data such as information on vital signs, biological measurements, sleep patterns, family histories and locations visited is being generated and collected at unprecedented rates by everything from fitness trackers to smartphone applications to online health calculators. With the prospect of combining this data with traditional health data sources for the purposes of health research, there is intense debate around who owns this data (i.e., user versus developer), who should be allowed access to it, and what purposes it can be used for. For studies conducted in academic institutions, research protocols are carefully reviewed by Research Ethics Boards, thus helping to ensure patient privacy is protected. However, commercial entities may be able to conduct big data research without similar oversight.

From an application perspective, since these data are not usually collected for health research purposes, how to handle missing data and possible input errors also pose challenges. Addressing these often requires making assumptions about the missing data, developing decision rules and algorithms, and using more advanced techniques (e.g., imputation). Furthermore, capitalizing on the vast amount of data also requires knowledge of innovative methods in data analytics and data science, for which uptake in health care is only beginning.

Although much progress has been made in reducing cardiovascular incidence and mortality in the last few decades, gaps in knowledge and cardiovascular care persist. Innovative methods from leveraging big data and using non-traditional data sources are gaining attention for their potential to contribute to ongoing advancements in cardiovascular health. While big data has generated much interest in the research community, its use is still in its infancy and how it can best be used to improve cardiovascular clinical care and patient outcomes has yet to be determined. Nonetheless, the use of big data in cardiovascular research has the potential to dramatically change how research is conducted in the future.

References

1. Tu JV, Chu A, Donovan LR et al. The Cardiovascular Health in Ambulatory Care Research Team (CANHEART): Using big data to measure and improve cardiovascular health and health care services. Circulation: Cardiovascular Quality and Outcomes 2015;8:204-212.
2. Tu JV, Chu A, Rezai MR et al. The Incidence of major cardiovascular events in immigrants to Ontario, Canada: The CANHEART Immigrant Study. Circulation 2015;132:1549-1559.
3. Tu JV, Chu A, Maclagan L et al. Regional variations in ambulatory care and incidence of cardiovascular events. Canadian Medical Association Journal 2017;189(13):E494-E501.
4. Farkouh ME, Domanski M, Sleeper LA et al. Strategies for multivessel revascularization in patients with diabetes. N Engl J Med 2012;367(25):2375-2384.
5. Yusuf S, Bosch J, Dagenais G et al. Cholesterol lowering in intermediate-risk persons without cardiovascular disease. N Engl J Med 2016;374(21):2021-2031.
6. James S, Rao SV, Granger CB. Registry-based randomized clinical trials- a new clinical trial paradigm. Nat Rev Cardiol 2015;12(5):312-316.
7. Lauer MS, D'Agostino RB. The randomized registry trial – the next disruptive technology in clinical research? N Engl J Med 2013;369(17):1579-1581.
8. Fröbert O, Lagerqvist B, Gudnason T et al. Thrombus Aspiration in ST-Elevation myocardial infarction in Scandinavia (TASTE trial). A multicenter, prospective, randomized, controlled clinical registry trial based on the Swedish angiography and angioplasty registry (SCAAR) platform. Study design and rationale. American heart journal 2010;160(6):1042-1048.

Appendix – Links to referenced websites

To learn more about the CArdiovascular HEalth in Ambulatory care Research Team (CANHEART) initiative, see http://www.canheart.ca.

To view data on cardiovascular indicators across Ontario, Canada's health service regions or Local Health Integration Networks, see http://www.canheart.ca/eatlas/.