The SPOR Canadian Data Platform: opportunity for multi-provincial research

Astrid Guttmann MDCM MSc

In April 2019, the Canadian Institutes of Health Research (CIHR) announced an $81-million funding commitment for an initiative called the Canadian Data Platform under the Strategy for Patient-Oriented Research (SPOR). The funding will enable, among other things, the development of a single portal through which researchers can request access to health administrative, demographic and social data from various federal and provincial sources. This comes 8 years after an international CIHR review panel called for "a Canada-wide effort to harmonize data sets and enable national linkages which would benefit all CIHR institutes and the Canadian research enterprise at large."[1] The ambitious proposal represents a welcome step forward to capitalize better on the rich population-based data assets in Canada. If successful, it will create a "distributed" network to facilitate multi-jurisdictional research, consolidate efforts to harmonize and validate data definitions, share technical and other expertise, and better support advanced analytics and infrastructure needed to realize the potential of an increasing array of "big data."[2]

Canada has the opportunity to be an international leader in population-based health and health systems research. The single-payer health systems capture whole populations tracked longitudinally through all publicly funded provincial health services. These administrative data, collected in the day-to-day work of the health system by clinicians (e.g., electronic health records) or government (e.g., physician billings), are the backbone for this research. Many provinces have health services research institutes, which both capitalize on these data and work closely with policy-makers to conduct applied research. In some provinces, health administrative data are linked to a wide variety of other service and demographic data (e.g., immigration, social services, education, child welfare). Manitoba has the most impressive array of linked intersectoral data, which have enabled world-class research on health and social equity.[3] These provincial data repositories are also the "backbone" for linkage with an increasing array of deeper clinical data. Examples include population-based laboratory data, electronic medical records and clinical-trial data, as well as research cohorts, which include imaging, genomic and other "omic" information. Although

KEY POINTS

- Canada has a rich array of provincial and federal health and social data assets, but important challenges exist for researchers to capitalize fully on these.
- The Canadian Institutes of Health Research has funded the development of a national data platform that will address some of the key barriers to multi-jurisdictional research.
- Further investment will be needed to ensure the platform goes beyond administrative data to include other "big data" (detailed clinical electronic health record and "omics" data) and support advanced analytics.

recent SPOR Support Unit funding has improved data access in provinces without dedicated data institutes, material differences in available data and access remain.

The overarching goal of the national data platform — which will be overseen by a consortium of leaders from provincial data centres, SPOR Support Units, the Canadian Institute for Health Information and Statistics Canada, working under the auspices of the Pan-Canadian Real-World Health Data Network[2] — is to catalyze more multi-jurisdictional research. All provinces and the Northwest Territories are represented, as well as knowledge users from provincial ministries of health and leaders in patient and public engagement and Indigenous health research.

Canada's 13 health systems provide a natural experiment for policy and program evaluation. Consider, for example, the ability to compare patient outcomes and system costs of differing pharmacare eligibility across provinces. To do so, researchers require access to comparable data to create similar patient cohorts and measure drug access or use and outcomes. Increasingly, disease-based quality improvement networks need multi-provincial data to evaluate interventions. As interest grows in both pragmatic trial design and linkage of dormant trials to administrative data for long-term outcomes,[4] access to comparable data is necessary for multi-provincial trials. Similarly, the ability of pan-Canadian cohorts such as the Canadian Partnership for Tomorrow to link to administrative data will enhance their value. Finally, multi-jurisdictional research allows pooling

of data for the study of uncommon outcomes. The Canadian Network for Observational Drug Effect Studies[5] is an example of a distributed data network that uses standardized protocols and meta-analysis across provinces to study real-world outcomes of drugs after they are marketed.

How will the national data platform help? At its most basic level, the platform will provide a single portal through which researchers will have better access to data from multiple jurisdictions, and dedicated personnel in every province and territory to help researchers navigate data holdings and processes to access data and conduct distributed analytics (or comparative analyses). This will overcome the well-known barriers to pooling data across provinces owing to provincial legislation. The platform will not change any of the current provincial requirements for data access, but may be a catalyst for evolution in those jurisdictions with more cumbersome processes for accessing data and fewer available data sets.

The platform will leverage existing and long-standing provincial and federal investments in health data and data infrastructure. Some provinces have well-developed capacity for data linkage and storage, and the platform will allow knowledge sharing with respect to a number of these technical areas as well as investment in the technology to create, share and fulfill data access requests and enable distributed analyses. Data science approaches to both data handling and analytics, including techniques such as natural language processing and machine learning or artificial intelligence, are important areas for collective work, as is the need for data validation.[6]

The platform proposes to validate disease definitions for a set of prioritized conditions. Diagnoses in physician billings are notoriously inaccurate and differences in available data across provinces pose challenges to standardizing definitions. For example, drug data (insulin and oral hypoglycemics) or laboratory data (glycosylated hemoglobin) can be used in some provinces, but not others, to identify cohorts with diabetes, so the ideal state of comparable cohorts across a number of conditions may not be realized. In addition, although electronic medical record data have the potential to contribute to standardizing disease definitions and provide rich clinical data, extraction costs even for structured data elements such as height and weight are substantial. Apart from cost, the technical challenges are immense with respect to extracting richer data from, for example, clinical notes, and in the context of a multitude of different electronic medical record vendor systems and data governance agreements. Finally, apart from disease definitions, researchers often develop algorithms and other definitions of constructs (e.g., primary care visits) for studies, and ensuring that researchers have access to "local knowledge" of the health system is critical to data validity and comparability.

Finally, there are many other important areas of leadership for this platform, including supporting Indigenous data sovereignty.[7] The platform could build on nascent efforts by some provincial research institutes to engage the public through public advisory councils in discussions on the use of large health data sets to drive research. Although the importance of and support for involving patients in health research is now well established, less work has been done in considering how the public should intersect with research organizations and researchers who use administrative data. Qualitative work in Ontario suggests the public is accepting of the use of their data for "the public good,"[8] but examples from both Canada[9] and England[10] indicate that communication and public engagement are critical to ensuring this research has public trust.

The SPOR Canadian Data Platform builds on existing CIHR–SPOR investments and provincial and federal data infrastructure in an ambitious effort to overcome well-known obstacles to conducting multi-jurisdiction research in Canada. The CIHR's substantial investment is a great step forward, but vaulting to international leadership will likely require additional investments to ensure the platform goes beyond administrative data to include other data sets (e.g., detailed clinical data in electronic health records and "omics" data), as well as to support advanced analytics including artificial intelligence and machine learning.

References

1. CIHR response and action plan — 2011 International Review Panel Recommendations. Ottawa: Canadian Institutes of Health Research; 2011. Available: www.cihr-irsc.gc.ca/e/44567.html (accessed 2019 Aug. 5).
2. Pan-Canadian Real-world Health Data Network (PRHDN) [Web page]. Toronto: PRHDN; 2019. Available: www.prhdn.ca (accessed 2019 Aug. 8).
3. Nickel NC, Chateau DG, Martens PJ, et al. Data resource profile: Pathways to Health and Social Equity for Children (PATHS Equity for Children). *Int J Epidemiol* 2014;43:1438-49.
4. Fitzpatrick T, Perrier L, Shakik S, et al. Assessment of long-term follow-up of randomized trial participants by linkage to routinely collected data: a scoping review and analysis. *JAMA Netw Open* 2018;1:e186019.
5. Suissa S, Henry D, Caetano P, et al. CNODES: the Canadian Network for Observational Drug Effect Studies. *Open Med* 2012;6:e134-40.
6. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015;12:e1001885.
7. Walker J, Lovett R, Kukutai T, et al. Indigenous health data and the path to healing. *Lancet* 2017;390:2022-3.
8. Paprica PA, de Melo MN, Schull MJ. Social licence and the general public's attitudes toward research based on linked administrative health data: a qualitative study. *CMAJ Open* 2019;7:E40-6.
9. Press J. Politics. Bank data furor threatens ability to compile accurate data: chief statistician. *The Canadian Press*; 2018 Nov. 1.
10. Presser L, Hruskova M, Rowbottom H, et al. Care.data and access to UK health records: patient privacy and public trust. *Technology Science* 2015 Aug. 11.