Digital research infrastructure to support federated computing on large scale biomedical datasets

Guillaume Bourque
McGill University

Michael Brudno[1]
University of Toronto
University Health Network

Steven JM Jones
Genome Sciences Centre, BC Cancer

Executive Summary
Data is dramatically transforming biomedical research and healthcare. Genomes now provide fundamental insight into our predisposition to diseases, response to therapies and environmental impacts on our health. Moreover, advances in artificial intelligence (AI) are revolutionizing diagnostic imaging and health record analysis, helping automate time-consuming, expensive workflows, and predicting interventions. As a global scientific community, our ability to interpret and utilize rapidly expanding genomic and health data is at a nascent stage, while the infrastructure needs are growing: by 2025 **>60 million patients will have their genomes sequenced** for healthcare purposes worldwide, requiring up to 40 exabytes of storage[1,2].

To fully realize the benefits of digital health, we need national-scale data and software resources with privacy policies and hardware infrastructure to securely store, share and interpret medical data using AI and other advanced analytic tools.

Growth of Health and Genomics Data
With today's rapid rate of digital and technological innovation, investment in digital health has seen significant expansion:

- The digital health global market will grow from $106B USD (2019) to >$639B USD by 2026[3].
- The digital genomics global market share in 2018 was $26.0B USD, and this is expected to grow to $50.4B USD by 2025, registering a 10% compound annual growth rate[4].
- Genome Canada has invested a total $3.9B CAD of funding towards 455 projects across various fields of genomic research[5].
- Digital healthcare-centric efforts have received substantial government funding in Canada (e.g. the Terry Fox Research Institute Marathon of Hope Cancer Centre Network and the Digital Health and Discovery Platforms (DHDP) were recently awarded $200 million of funding)[6].

This mirrors a rapid expansion in the amount of health-related digital data available to researchers and clinicians. The widespread use of digital technology, collection of large-scale clinical datasets,

[1] Contact author: brudno@cs.toronto.edu

and increasing availability of cost-efficient, rapid sequencing techniques has made digital health incredibly accessible, and it is quickly becoming a staple component of healthcare practices.

Current Challenges in the Field of Healthcare Data

Collaboration across clinical, research and industry sites is necessary for genomic and health data to be used effectively. Current Canadian infrastructure, however, is not conducive to efficient data storing, sharing or analysis, leaving data often fragmented, siloed and inaccessible and impeding collaboration and scientific insights that could be gleaned from the diverse Canadian population.

- The **massive volume of sensitive health data** generated today requires software and computer technology beyond what is available at individual centers across Canada.
- **Critical mass of patient data** is needed for advanced AI approaches, which often relies on the sharing of consented data between scientific and medical centers.
- Current Compute Canada and Canadian private sector **resources are inadequate** for the complexities of hosting, analyzing, protecting and sharing this data at the needed scale.
- **Many data systems are developed for specific purposes** only, with minimal design considerations into how information may be used for additional or future means[7].
- Mechanisms and technologies that **facilitate greater interoperability** of tools and services to enable effective sharing of information at distributed sites across Canada are lacking.

Canadian universities, hospitals and research centres need sufficient privacy, software, and computational infrastructure to facilitate novel information sharing modalities and to reap the full benefits of Canadian health data. This includes collaborative computing environments to run complex data mining and analysis operations across petabytes of public and private genome sequences and their associated clinical information. Infrastructure should prioritize secure data storage, sharing and analysis, as well as interoperability between sites. Together this will place Canada as a leader in health research internationally.

Health Data And Infrastructure Platforms Across Canada

Integration of diverse types of clinical and -omics data through the sharing of large datasets across the scientific community is fundamental to driving scientific and clinical progress. Coalescing information from datasets collected at multiple institutions allows researchers and clinicians to study health and disease-related patterns across diverse groups of people on an unprecedented level. In Canada, several efforts to do this are present:

- **All For One:** Rare disease patient numbers are often insufficient at any one site to drive research progress. Genome Canada's All for One is a pan-Canadian initiative giving clinicians and researchers access to national-scale data to forward rare disease research and patient care[8].
- **Canadian Distributed Infrastructure for Genomics (CanDIG):** A platform connecting McGill University, Hospital for Sick Children (SickKids), the University Health Network (UHN), Michael Smith Genome Sciences Centre, Jewish General Hospital and Université de Sherbrooke. Sites can access and analyze an array of human genomics data, encouraging national collaboration in health data projects[9].

- **Terry Fox Research Institute Digital Health and Discovery Platform (DHDP):** A Canadian platform that facilitates data presentation and accessibility for real-world evidence-based healthcare discoveries[10]. CanDIG is one of the initiatives enabling the DHDP.

Initiatives such as these must be cognisant of the sensitive nature of clinical and biomedical data, and a Canadian national-scale digital research infrastructure (DRI) for the use of such data must be developed with careful consideration. Addressing the challenges outlined above through a harmonized pan-Canadian DRI will enable effective and responsible sharing of health data and catalyze national projects that drive and demonstrate the value of data sharing, including new features such as radiomics, standardized outcomes, pathomics, and other emerging data types.

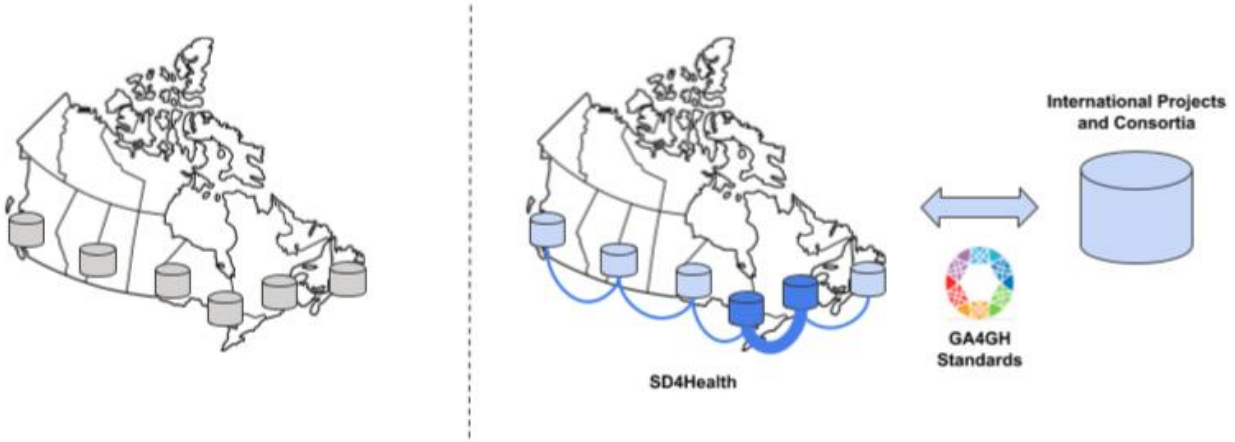

Figure 1: Currently, genomic and health datasets generated in Canada are often in silos, limiting data sharing and analysis (left). Programs such as HPC4Health and SD4Health create a secure pan-Canadian platform for advanced data analysis to fully exploit these valuable datasets (right). They contribute and build upon the GA4GH's interoperability standards, enabling collaboration between Canadian and international researchers.

Canadian hardware infrastructure that can be built upon for these initiatives include HPC4Health and SD4Health, two health computing nodes working together to bring high performance computing (HPC) to researchers with sensitive datasets in Ontario and Quebec, respectively.

- **HPC4Health**: A private cloud service that provides researchers and clinicians from Ontario hospitals (SickKids, UHN, CHEO, ICES, others) with cloud-based HPC that is secure and satisfies personal health information privacy requirements. HPC4Health infrastructure policies address aspects of both security and privacy, including account & passwords, backup & retention, change management, encryption, data de-identification, and incident response.
- **SD4Health:** A platform that connects 7 of the largest research institutions in Québec. Similar to HPC4Health in Ontario, it provides unified security solutions for health research projects. Participating sites can analyze and share datasets nationally and internationally via Global Alliance for Genomics and Health (GA4GH) protocols and have access to a suite of innovative AI tools for the analysis of health data.

The collaborative work between SD4Health and HPC4Health, including the sharing of policies and infrastructure across both sites is an exemplar for building pan-Canadian federated infrastructure

that can be leveraged by scientists across the country. Together with BC Cancer Genome Sciences Centre (BCGSC), our teams are working to develop common application programming interfaces (APIs) that enable data access at all three institutions through internationally leading projects.

Data Federation To Enable Canadian Healthcare DRI

To unite expertise and data access across Canadian hospitals in various healthcare jurisdictions, a national federated approach is needed. This will enable the querying and analyses of national-scale genomics and human health datasets while the data itself remains securely and privately controlled. Policies, governance and architecture, APIs and ontologies tie all participating sites together into a single coherent network. This includes:

- **Foundational policies, governance and technical architecture** that unite privacy, security and authentication practices across sites, such that provincial policies are followed while data access and use is maximized. Industry-standard security frameworks block intrusion, authenticate users, roles, and digitally signed applications.
- **Standardized APIs** that unify access to data collected and siloed by distinct systems[7]. APIs are used by researchers to query and analyze distributed datasets and by data stewards to preside over the data they have been entrusted with.
- **Standardized ontologies** that formally define shared vocabularies and harmonize data coming from different systems by providing a common way to annotate and classify diverse datasets. Medical consent ontologies prevent violations by describing informed consent, opt-in, opt-out, terms-of-use, specific research usage etc.
- **Improved accessibility** for patients to digitally consent, multilingual and multicultural support and dynamic consent practices that allow patients to control how their data is used.

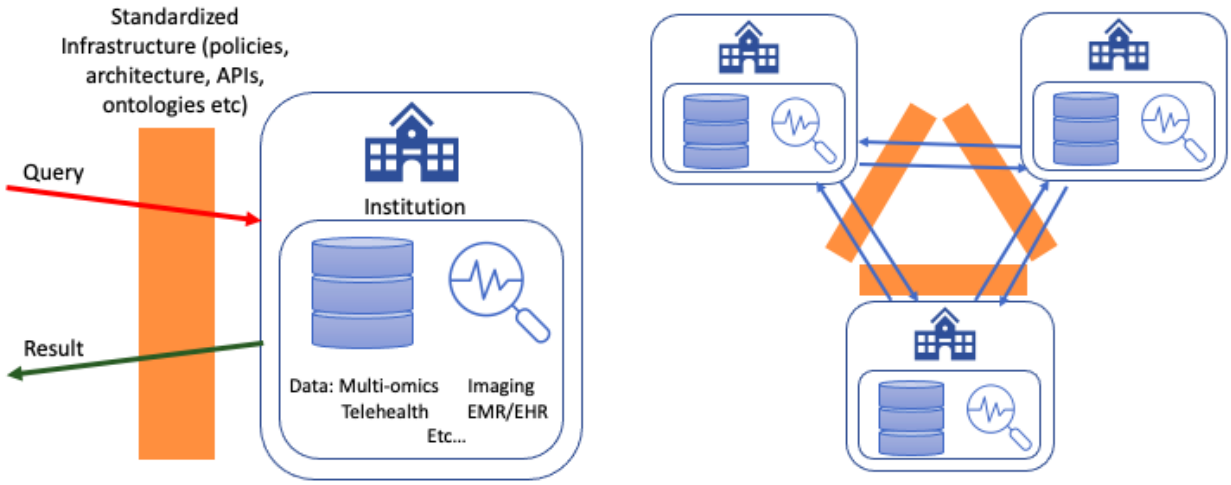

Figure 2: A federated approach to data sharing involves standardized infrastructure through which researchers can access Canada's national wealth of digital health information seamlessly from within their own workspaces and institutions. A national network of hospitals and medical centers connected through a federated platform will comply with provincial privacy regulations while contributing to the diversity of data available to Canadian researchers.

A scalable, federated platform that leverages these points to enable collaboration between Canadian researchers and clinicians will advance research, medicine and patient care. Dozens of individual systems exist in each research hospital, and connecting them within and between institutions across Canada will be revolutionary. An open-source stack of standards-led tools is needed to integrate institutional research/clinical databases (e.g., pathology, biobanking, imaging, genomics, medical data, etc.) to support semantic findability, accessibility, and interoperability. Consent practices need to empower patients, motivating them to contribute their data. These technological innovations will let networked researchers work collaboratively with each other and with members of the biomedical industry (pharma, biotech, medical device and imaging companies).

Canadian DRI Efforts in an International Context

The number and diversity of uncontrolled variables that impact our health is exceedingly large, and the scientific community must cast a wider net in the collection of health-related data to fully understand how certain conditions and their treatments affect us. The progressive digitization of healthcare will be a powerful enabler for this, allowing researchers to collectively amass data from large and heterogeneous populations. Canada is ideally positioned for this ambitious development.

- Our **broad diversity** can capture nearly the full range of global variation in biomarkers, genetic determinants for disease, responses to treatment and more.
- Canada is a **world leader** in genomics, machine learning, bioinformatics and medical research.
- National **efforts are already in place** and can be built upon (e.g. CIHI, CanDIG, DHDP).

Canada needs a national infrastructure for the dissemination and use of health data to accelerate our progress in biomedical research and healthcare, and enable national and international collaboration. CanDIG and DHDP take a federated approach to this, where geographically and jurisdictionally distinct sites are connected with common principles that prioritize data privacy and security. They implement standards of practice founded by the GA4GH, which aims to establish policy frameworks for responsible and effective health-related data sharing worldwide. Similarly, All for One utilizes a policy toolkit for the development of interoperable and standardized resources that protect patients and meet the needs of different clinical and research sites. Yet these initiatives have separate web portals organized in varying ways, making it difficult to get an overview of available datasets across the whole of Canada.

Importantly, to remain internationally competitive, Canadians must embrace our strengths through access and sharing of our collective resources. In addition to providing Canadian researchers robust and diverse datasets to study the predisposition, onset, progression and treatment of disease, a national healthcare data-oriented DRI will ease international research collaborations, particularly if GA4GH standards are applied. GA4GH has developed open standards for APIs, consent forms, ontology, repositories and more, which have all been implemented by >600 research and healthcare organizations and projects worldwide[2]. A federated national DRI that likewise implements GA4GH standards will give Canadian researchers the tools needed to access, use and contribute to large international datasets, improving the quality of not only our own scientific community, but of research across the globe.

References

1. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLOS Biology* **13**, e1002195 (2015).
2. Birney, E., Vamathevan, J. & Goodhand, P. Genomics in healthcare: GA4GH looks to 2022. *bioRxiv* 203554 (2017) doi:10.1101/203554.
3. Digital Health Market Share Trends 2020-2026 Growth Report. *Global Market Insights, Inc.* https://www.gminsights.com/industry-analysis/digital-health-market.
4. Digital Genome Market Global Projections 2019-2025 Report. *Global Market Insights, Inc.* https://www.gminsights.com/industry-analysis/digital-genome-market.
5. Genome Canada - Annual Report 2019-2020.
6. Impact Report 2018/2019. (2020).
7. Hermann, A. Federated Data Systems: Balancing Innovation and Trust in the Use of Sensitive Data.
8. All for One Policy Toolkit | Genome Canada. https://www.genomecanada.ca/en/all-one-policy-toolkit.
9. CanDIG. https://www.distributedgenomics.ca/.
10. Digital Health and Discovery Platform. https://www.tfri.ca/our-research/digital-health-and-discovery-platform.