Building a data science platform for better health

A White Paper submitted to the New Digital Research Infrastructure Organization (NDRIO)

December 14, 2020

Submitted by:

Laura C. Rosella[1], Ajay Agrawal, Phillip Awadalla, Adalsteinn Brown, Jeffrey Brook, Lori Diemert, France Gagnon, Marzyeh Ghassemi, Avi Goldfarb, Rick Glazier, Michelle Greiver, Alan Katz, Kathy Kornas, Douglas Manuel, Kimberlyn McGrail, P. Alison Paprica, Fahad Razak, Michael J. Schull, Charles Victor, Amol Verma, Sabrina Wong

BACKGROUND AND RATIONALE

There are few aspects of society that are more complex than health and health care. The complexity is driven by strong sociological, economic, environmental and biological forces that affect systems and individuals. While this complexity has been well recognized[1], Canada's ability to study this using data-driven approaches with the most advanced analytic methods is lagging among peer nations. The future digital infrastructure must allow for the integration of research data spanning complex and varied datasets that collectively capture patient and population experiences within and outside the health system (i.e., clinical, health services utilization, social determinants, and environmental data), along with specialized software and analytical tools to effectively utilize these newly integrated resources for data-driven research. Infrastructure must strive for unified representations of data that incorporate the full spectrum of the determinants of health using data from multiple sources versus advancing siloed data within a homogenous data system. Indeed this has been described as the "holy grail" of deep learning research[2] that is currently lacking in Canadian digital research infrastructure (DRI).[3] The collective power of this resource will far exceed that of its individual pieces and will create possibilities for new research innovation, which is currently not possible. This vision requires a level of collaboration rarely seen in the Canadian health data landscape.

The ideas that are included in this white paper represent a multidisciplinary team of scientists who have collaborated to conceive and design a unique health data infrastructure for Canada. Since 2019, the authors of this paper have been working together with the goal of creating a health DRI that spans multiple health and health-related data sources, and this paper presents learnings and recommendations that emerged through this collaboration. We focus on infrastructure to enable Artificial Intelligence (AI), Machine Learning (ML) and other computationally intensive analytics as we see this as the largest gap and the most relevant for the future DRI state. The urgency of these needs was also recently highlighted in the 2020 AI for Health task force report.[3] Specifically, we focus on the DRI needs to produce world-leading research in three key areas: (1) Prediction (i.e., creating rich clinical, administrative and population data on patients and populations for developing AI/ML prediction tools for health); (2) Health system and population health deployment (i.e., building infrastructure that will allow latest advances in advanced analytics to support a range of health care delivery and prevention research); (3) Discovery research (i.e., computational and statistical methods development, creating large accessible datasets and enabling the use of new variables and measures from newly integrated complex data sources). This infrastructure must also bring together people to forge new collaborations between subject matter experts and experts in advanced analytics. Importantly, these efforts must be underpinned by strong governance of personal health information and efforts to build capacity for health system decision-makers to be able to use

[1] Designated contact person, laura.rosella@utoronto.ca, other authors listed in alphabetical order according to last name

the insights generated. This combination of top analytic talent and expertise in accessing, governing, and managing these data fosters the ideas needed for the development, testing and deployment of new data science approaches for health across Canada.[4] This includes predictive analytics, which we believe is more than building algorithms; it also includes building the prediction to the implementation pipeline that reflects varied data and application environments. Future DRI infrastructure must evolve from current infrastructure by enabling an integrated, streamlined and standardized predictive research workflow for health data science. This would produce methodological breakthroughs that will lead to innovations to improve population health, drive better care and efficiencies in Canada's health care system, that can be learned around the world.

CURRENT ISSUES

Canada is lagging internationally with respect to investigations related to health data. For example, the European Union EHR4CR project has received €16M ($25M CAD) to develop solutions to use electronic health record (EHR) data for clinical research in 11 hospital sites across 5 countries. In the US, the National Patient-Centered Clinical Research Network (PCORnet)[5] is investing more than US$70M ($90M CAD) to integrate data from EHRs for research. Despite these significant investments in our peer nations, none have tackled the breadth of data needed for innovative research applications.

Canada has tremendous data advantage potential stemming from large investments in population-level data collection, single-payer health systems covering a diverse population, and the ability to integrate across public health and clinical health systems. However, our current health data infrastructure is not internationally competitive, particularly to accommodate advanced data science applications and harness investments from existing data sources. This failure has been confirmed through many national expert advisory panels.[3,6-8] Furthermore, AI/ML are growing areas of international focus for its potential to change society, yet Canada's research productivity in the AI health sciences sector is currently lagging behind the USA and Europe.[9] By building DRI dedicated to bringing together varied data sources and expertise to enable advanced data science in a high computing environment for health information, Canada has the potential to build the most advanced health data infrastructure on the most diverse population data structures in the world. We envision an infrastructure that would be more comprehensive than even the most sophisticated national population-based data systems established elsewhere in the world, such as the HDR-UK[10,11], UK Biobank[12], Integrated Data Infrastructure in New Zealand[13], and the PCORnet in the USA.[5] The resulting development of open software deployment tools will also address gaps in implementation and deployment in the health system.

The envisioned DRI would also significantly expand Canada's ability to attract world-class talent, filling the big data analytics gap identified by Canada's Big Data Consortium.[14] To attract the world's AI talent, Canada needs an infrastructure that can enable research-ready population health data that incorporates the complexity for integration across domains. Population-wide data infrastructure will be a major draw for trainees from within and outside of Canada, with international trainees who are interested in coming to Canada to work with globally unique population health data. Without this infrastructure, Canadian health data science researchers cannot remain competitive internationally, putting us at risk of losing existing talent who are unable to carry out their research because of infrastructure gaps.

DOMAINS NEEDED FOR FUTURE DIGITAL INFRASTRUCTURE

We believe there is an urgent need for health DRI in Canada that can enable new data science research in the health domain that will impact the health of populations and health systems. We propose the future DRI must advance current health infrastructure limitations in several important ways:

(1) Canadian health-related datasets are held in separate environments and/or on platforms that lack the computing infrastructure for computationally intensive analytics. The next breakthroughs in health data science will be achieved by the **integration of comprehensive data sources that capture the full determinants of health on a computing infrastructure that allows for advanced analytics.**

(2) Unlike most health data infrastructure that exist provincially in Canada, new health DRI must **enable pan-Canadian research initiatives**. Multi-province data initiatives do exist in Canada; however, these are generally

focused on a specific health outcome, data type (i.e. only health administrative data in the case of HDRN) or health care setting (i.e. primary care).[15] Digital infrastructure must leverage the foundational work of the HDRN, which is focused on cross-provincial research using harmonized data, algorithms, and protocols for administrative data, and build on this by integrating other national data (CanPath[16], CANUE[17], CPCSSN[18]). Furthermore, strong within provincial initiatives exist, such as General Internal Medicine Inpatient Initiative (GEMINI), which has impressively demonstrated how Canada's complex and varied clinical data streams in hospitals, including laboratory, medications, and imaging, can be extracted, aggregated and used for analytics.[19] Infrastructure must be built for national access across Canada, enabled by provincial high-performance data trust environments that can be scaled in every province. To this end, we support the individuals and organizations that are submitting White Papers to NDRIO that focus on the components and enablers of this integrated infrastructure: Health Data Research Network Canada (HDRN Canada, P. Alison Paprica), Diabetes Action Canada (Gary Lewis), the Canadian Primary Care Sentinel Surveillance Network (CPCSSN, Sabrina Wong), and the Ontario Primary Care Learning (ORACLE) Network (Michelle Greiver).

(3) Existing health data infrastructure fails to address barriers for use, such as high-performance remote access, and as a result, these data are grossly underused by AI/ML researchers. In other words, we have AI/ML potential in Canada through increasing efforts of talent recruitment but lack the infrastructure to enable this talent to have research and system impact. **Infrastructure must focus on practical uptake and acceleration of AI/ML approaches for multimodal integrated health data by building high performance capabilities both within and across provinces**. Further, health data infrastructure in Canada has not been co-created with AI/ML researchers, resulting in retrofitted analytic systems. Critical to infrastructure development is the co-creation of DRI with leading data science and AI experts and health system experts who can appropriately inform on technical specifications.

(4) There are few health data initiatives in Canada that includes **software development to enable deployment of the analytic outputs into the health system**, which has resulted in the underuse of predictive models in practice. Future DRI must address this concern by including open-source predictive algorithm deployment that will flexibly enable the deployment of AI/ML algorithms on varied platforms that exist in health settings.

FUTURE DRI FOR POPULATION HEALTH

We envision infrastructure that will involve unprecedented applications and technical requirements, along with highly specialized personnel who will design, develop, build data resources, software and analytical tools, appropriate computing needs and create the foundation for deploying AI/ML innovations. We have identified the critical components of future DRI for health in Canada to include:

- **High-Performance Data Trust Environments:** *Infrastructure and personnel (data managers and system administrators) required to support the integration of large-scale data, spanning the population's experience within and outside of health settings.* This includes structured and unstructured data that capture socioeconomic and demographic, clinical (e.g., health utilization data, electronic medical record (EMR)/EHR data with free text), environmental (e.g., satellite, aerial and ground-based imaging, walkability, greenspace, individual smartphone data streams capturing geographic locations and corresponding environmental factors) and biological dimensions (e.g., genetic information, biomarkers). These Data environments must meet essential requirements for their timely, fair, safe and equitable use.[21]
- **Data Architecture and Pipelines:** *Re-usable and extensible infrastructure that supports the data-to-analysis pipeline required for AI/ML and other computationally intensive analytics.*
- **Predictive Analytics Deployment:** *Software and developers to support the development and implementation of predictive analytics.* Software components would enable the development, data collection and evaluation of the real-world use of the predictive analytics/AI and decision tools across geographies and institutions as well as for providers and patients.

Governance

NDRIO has a role in influencing a modernized organizational structure and governance model to enable the infrastructure capabilities we have proposed for future DRI. The governance structure should include representatives from all collaborating institutions, as well as advisory committees, which include the public, patients, clinicians, scientists, health system decision-makers and other end-users. We envision several domains to form a synergistic data governance and management approach:

- **Data Acquisition and Preparation:** To oversee the application of pre-specified data regulations for integration and use.
- **Data Access:** To establish rigorous yet feasible data access guidelines and procedures.
- **Stakeholder and Public Engagement:** To establish processes and timelines to engage stakeholders, including clinicians, public health, patients, communities, health administrators and governments.
- **Capacity Building:** To focus on engaging end-users as well as support sustainability and growth.

BENEFITS AND PATHWAYS FOR KNOWLEDGE TRANSFER

Canada's new DRI must catalyze new research collaborations between siloed areas that have traditionally not worked together, namely between public health, medicine, environmental, engineering and computational sciences. These interdisciplinary collaborations will develop new approaches for using data to transform health systems, improve population health and lead to improved care, which could benefit every country in the world. Population-based multi-component infrastructure will provide unprecedented opportunities for interdisciplinary training and collaboration that bridges population health and patient care, ensuring that Canada is well-positioned to address complex health system challenges through innovative data science approaches.

Training, skilled workforce and commercial applications

Data within the described DRI would provide researchers and trainees access to new data resources for training, including open-access datasets that comply with privacy legislation. Long-term, this will support capacity within Canada's skilled workforce to apply advanced health technologies and data-driven strategies. Through the integration of multiple data streams, this DRI will create unprecedented scholarly opportunities for multidisciplinary collaborations between researchers from the computational and health sciences. Currently, while Canada is aggressively pursuing AI/ML talent, Canada lacks a sufficiently skilled workforce to apply advanced health technologies through data-driven strategies and advanced analytics. The talent gap in Canada, identified by Canada's Big Data Consortium,[14] was estimated between 10,500 and 19,000 professionals with deep data and analytical skills in 2014. This skills gap is particularly large in the field of population health and health systems. A new health DRI will enable the necessary high-powered computing and use of cloud data to support national access to datasets, which is integral to building skills and capacity for health AI research that will benefit all Canadians. The infrastructure will be a major draw for researchers and students to come to Canada by providing them with, amongst other things, unparalleled data on a diverse population across sectors, with high performance computing for new and emerging analytics. This will have major implications for increasing the supply of AI talent and training new students from diverse domains such as health policy, epidemiology, environment, clinical medicine, and computational sciences. Built in from the start, we must include pathways for commercialization, respect legal and regulatory constraints, and ensure regulations are transparent, ethical and aligned with the values of Canadians. Relationships with university-based accelerator programs, such as UofT's Rotman School of Management's Creative Destruction Lab, a program for seed-stage, science-based start-ups, are needed to support opportunities for innovative commercial applications, which are particularly useful avenues for research-driven applications that can be scaled.

Health and economic benefits of the future DRI

Health care costs are rising steeply around the world. Globally, Canada has among the highest health care spending with total costs for 2018 estimated at $253.5 billion or $6,839 per Canadian, representing 11.3% of Canada's gross domestic product.[21] The societal benefits from Canada's data investments are not being fully realized given our health system under-performance compared to other countries.7 Canada's health care system continues to lag in innovative advancements - at a great cost to society in the form of suboptimal health care delivery and inefficient care. Our AI-based breakthroughs on the use of data are critical to the transformation of health care systems and to improving care, management and treatment of patients. **Future DRI in Canada must focus on integrating complex exposure histories from the social, clinical, environmental and biological dimensions, as well as multiple disease and health system outcomes through linkage with a wide range of structured and unstructured data.** This will enable further innovation in data generation and the use of new approaches, such as unsupervised learning methods, which will reveal new insights into the complex nature of health conditions and health in society. The algorithm deployment tools that would be developed will support a multitude of predictive algorithms for use by clinicians and decision-makers. The potential of this work to transform the function and planning of health systems is on the grandest of scales. The result would be a paradigm shift that transforms the way we research health and deliver health care in Canada and the world.

NDRIO AND THE FUTURE OF HEALTH DRI

There have been limited large-scale efforts to support infrastructure that enables advanced analytics and data science approaches for population-level data, which span the patient experience, within and outside health settings, and which can be **scaled nationally.** NDRIO is encouraged to acknowledge a need for a health DRI that (1) integrates currently siloed health, environment and population data sources, (2) enables the use of this data with advanced AI/ML analytics, which has additional computational needs, and (3) includes software and developers to support the implementation of the resulting predictive analytic tools into health settings. As a community practice leader, NDRIO can host national discussions with researchers, governments, and regulators related to leveraging current distributed and integrated health data models and developing AI/ML capacity in other provinces through remote access to create such a system. In addition, NDRIO can bring together academic and industry partners to form collaborations that will support the deployment of predictive tools that would result from population infrastructure. Specifically, in order to support the DRI of the future, NDRIO can ensure that efforts to integrate across data sources capturing the full spectrum of the determinants of health, avoid being hindered by provincial or institutional silos and is co-created with experts in AI/ML and other computational analytics to ensure that its technical specification matches the needs of contemporary analytics.

Canada’s *opportunity* is the combination of both leading researchers, and availability of large population-based data assets, contrasting countries like the US where much data is privately held within hospitals and insurance organizations. A unique aspect of building this capacity in Canada is the incredible multi-cultural society we represent. This innovative health infrastructure will enable research to be generated on data from the most diverse population in the world, which is a strength for the transferability of evidence and technology across jurisdictions and will have implications globally. The newly built infrastructure will support a wide range of advanced data science approaches for health that are currently not possible. Without this infrastructure, Canadian health data science researchers cannot remain competitive internationally, and Canada’s health care system will not benefit from innovative AI/ML-based breakthroughs. We envision a population health infrastructure that will focus on developing algorithm deployment tools. The breakthroughs resulting from the proposed infrastructure will support prediction, implementation and discovery research, generating insights relevant to the health and well-being of our population, and contribute to new ways to support health system transformation.

NDRIO has a role in ensuring that the resulting infrastructure bridges population health and patient care platforms, as well as create technology capable of accelerating the cycle of assessment and subsequent improvement in health management, environment and population well-being. Our pan-Canadian perspectives build from successful independent initiatives and multidisciplinary expertise. The combination of data assets, high-performance data trust environments, and predictive analytics deployment would solidify Canada’s position as a global leader in health data science, generate insights relevant to the health and well-being of our population, and deliver significant health and economic benefits to Canadians.

References

1. McMurty A. Reinterpreting interdisciplinary health teams from a complexity science perspective. *Univ Alberta Health Sci J.* 2007;4(1):33-42.
2. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *Ieee Journal of Biomedical and Health Informatics.* 2018;22(5):1589-1604.
3. Task Force on Artificial Intelligence for Health (AI4Health). Building a Learning Health System for Canadians. Toronto ON: Canadian Institute for Advanced Research; 2020. Link
4. McGrail K, Jones K, Akbari A, et al. A position statement on population data science. *International Journal of Population Data Science.* 2018;3(1).
5. Consortium PP, Daugherty SE, Wahba S, et al. Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research. *Journal of the American Medical Informatics Association.* 2014;21(4):583-586.
6. Council of Canadian Academies. Accessing Health and Health-Related Data in Canada. Ottawa ON: The Expert Panel on Timely Access to Health and Social Data for Health Research and Health System Innovation, Council of Canadian Academies; 2015. Link
7. Naylor D, Fraser, N., Girard, F., et al, . *Unleashing Innovation: Excellent healthcare for Canada - Report of the Advisory Panel on Healthcare Innovation.* Ottawa, ON: Her Majest the Queen in Right of Canada;2015.
8. Ontario Introduces 24 Ontario Health Teams Across the Province to Provide Better Connected Care: Ontario Health Teams Part of Province’s Plan to End Hallway Health Care. *New Release*2019.
9. Shoham Y, Perrault R, Brynjolfsson E, et al. *The AI Index 2018 Annual Report, AI Index Steering Committe, Human Centered AI Initiative.* Stanford, CA2018.
10. Naghavi M, Makela S, Foreman K, O'Brien J, Pourmalek F, Lozano R. Algorithms for enhancing public health utility of national causes-of-death data. *Population Health Metrics.* 2010;8(1):9.
11. Sebire NJ, Cake C, Morris AD. HDR UK supporting mobilising computable biomedical knowledge in the UK. *BMJ Health & Care Informatics.* 2020;27(2).
12. Boyd JH, Ferrante AM, O’Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC health services research.* 2012;12(1):480.
13. Atkinson J, Blakely T. New Zealand’s Integrated Data Infrastructure (IDI): Value to date and future opportunities. *International Journal of Population Data Science.* 2017;1(1).
14. Canada's Big Data Consortium. *Closing Canada's Big Data Talent Gap.* Toronto, ON2015.
15. Butler AL, Smith M, Jones W, et al. Multi-province epidemiological research using linked administrative data: a case study from Canada. *International Journal of Population Data Science.* 2018;3(3).
16. Dummer TJB, Awadalla P, Boileau C, et al. The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention. *Canadian Medical Association Journal.* 2018;190(23):E710-E717.
17. Brook JR, Setton EM, Seed E, et al. The Canadian Urban Environmental Health Research Consortium – a protocol for building a national environmental exposure data platform for integrated analyses of urban form and health. *BMC Public Health.* 2018;18(1):114.
18. Birtwhistle R, Keshavjee K, Lambert-Lanning A, et al. Building a Pan-Canadian Primary Care Sentinel Surveillance Network: Initial Development and Moving Forward. *The Journal of the American Board of Family Medicine.* 2009;22(4):412-422.
19. Verma AA, Guo Y, Kwan JL, et al. Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. *CMAJ open.* 2017;5(4):E842.
20. Paprica PA, Sutherland, E., Smith, A. et al. Essential Requirements for Establishing and Operating Data Trusts: Practical Guidance Based on A Working Meeting of Fifteen Canadian Organizations and Initiatives. *International Journal of Population Data Science.* 2020;5(1).
21. Canadian Institute for Health Information. *National Health Expenditure Trends, 1975 to 2018.* Ottawa, ON 2018.