A Standards-Based Digital Infrastructure for Secure Sharing of Human Biomedical Research Data

Authors: Lincoln Stein, Christina Yung, Philip Awadalla, Christine Williams, and Laszlo Radvanyi, Ontario Institute for Cancer Research (OICR)
Trevor Pugh, OICR and Princess Margaret Cancer Centre
Jeremy Adams and Peter Goodhand, Global Alliance for Genomics in Health
Date: 14 December 2020

Contact: Lincoln Stein <lincoln.stein@oicr.on.ca>

To translate Canada's leadership in big data analytics and artificial intelligence into tangible advances in precision medicine, we must break down the barriers for responsibly sharing human biomedical research data by building a national digital infrastructure for storing, sharing, integrating and analyzing large biomedical research data sets under a responsive and responsible data governance regime that balances human subject's privacy needs against the imperative to make collected data as widely available to the scientific research and clinical communities as possible.

The Problem

Canada is blessed with a robust collection of high performance compute clusters, high-speed provincial and national networks, and vibrant computer science, machine learning and biological research communities. Artificial intelligence (AI) and machine learning technologies are advancing at a breakneck pace, and seem poised to revolutionize both clinical medicine and biomedical research. In the near future, computer algorithms will help doctors and public health officials identify disease at its earliest stages, assist in the design of new generations of highly effective drugs, and provide doctors with precision medicine tools that match a patient's clinical, molecular and radiological profiles with the course of therapy most likely to be effective.

A major impediment to achieving this goal are the numerous practical, social, and legal impediments to sharing knowledge and data among researchers from different disciplines. Our systems for exchanging scientific data were established in the mid 20th century and are now creaking under the strain of the huge data sets that are routinely collected by clinical and biomedical researchers. Examples of key information-rich data sets collected in human health and disease research include high-throughput genome sequencing, single-cell expression profiling, digital radiology and tissue imaging. Data sets can involve thousands to hundreds of thousands of participants who have given their consent to contribute to scientific research, and can be mind-bogglingly large (a typical human genome takes as much space to store as a thousand feature-length 4K movies). However, our data-sharing procedures date to the days when the whole data set could be included in a single table printed in a paper publication.

In recent years, the biomedical research community has reached a consensus on data sharing called the FAIR principles,[1] for Findable, Accessible, Interoperable and

Reproducible. However, many practical obstacles to responsible data sharing remain:

- *No place to store the data.* These are large data sets. For example, a typical human genome equivalent to streaming a series of 4K HD movies continuously for a period of 2 months. While there are a handful of repositories for genomics and biomedical imaging data located in Europe and the United States, these are themselves straining under the pressure of international demand. It takes significant time and effort to submit a data set to one of these repositories, in some cases taking years from submission to acceptance. Some newer types of data have no usable repository at all, and in any event there are no biomedical data repositories of sufficient capacity located on Canadian soil.

- *Outdated processes for reviewing and granting access to shared data sets.* Data sets from human studies often contain potentially identifying information such as unique patterns of genetic variants, and for this reason access to such data sets is mediated by governance structures including institutional Research Ethics Boards (REBs) and Data Access Committees (DACs) jointly responsible for ensuring that the confidentiality of human subjects' data is preserved and that the proposed use of the data is consistent with the subjects' informed consent. However, these mechanisms are highly manual, effortful, and do not scale well when access is needed to the data produced by more than one research study. As a result, many innovative ideas that depend on data integration are stymied.

- *Data is stored distant from where it is needed.* Before using the data hosted by most existing biomedical data repositories, researchers have to first download the data to their local laboratory computers. However, these data files can be extremely large, and many smaller laboratories lack the resources to hold copies of the data set, rendering the data effectively inaccessible.

- *No standards for data generation, submission, search or access.* Most existing repositories of biomedical data use *ad hoc* methods for submitting, describing and retrieving data objects. This creates interoperability roadblocks and hinders the creation of general purpose query and retrieval tools.

The Solution

We propose the creation of an open source software suite for submitting, storing, finding, integrating, and responsibly sharing the fruits of Canadian biomedical research (**Figure 1)**. Recognizing that research communities have different needs, the system will be designed to provide research communities with the building blocks needed to create the data archiving, integration and exchange hubs that best fit their needs. The services to be provided include uploading and validating data types, registering projects, user access management, and handling multiple data governance models**.**

For example, a community of researchers studying variations in the genome responsible for rare inherited diseases could congregate around a repository in which the central data objects are family pedigrees, phenotypes and genomic variants. Access to the data would be open to any researcher whose institution's research ethics review

board has certified them for genetic disease research. In contrast, a community whose goal is to develop machine learning systems to use anonymized chest X-rays and clinical data to identify patients who will develop severe COVID-19 could create a repository specialized for handling imaging data, run on a cloud compute system with access to GPUs for deep learning, and be open to any member of the public. Because the software suite will handle these generic data sharing services, research communities can focus on the unique needs of their projects, and funding agencies do not end up paying multiple times for project-specific data sharing systems.

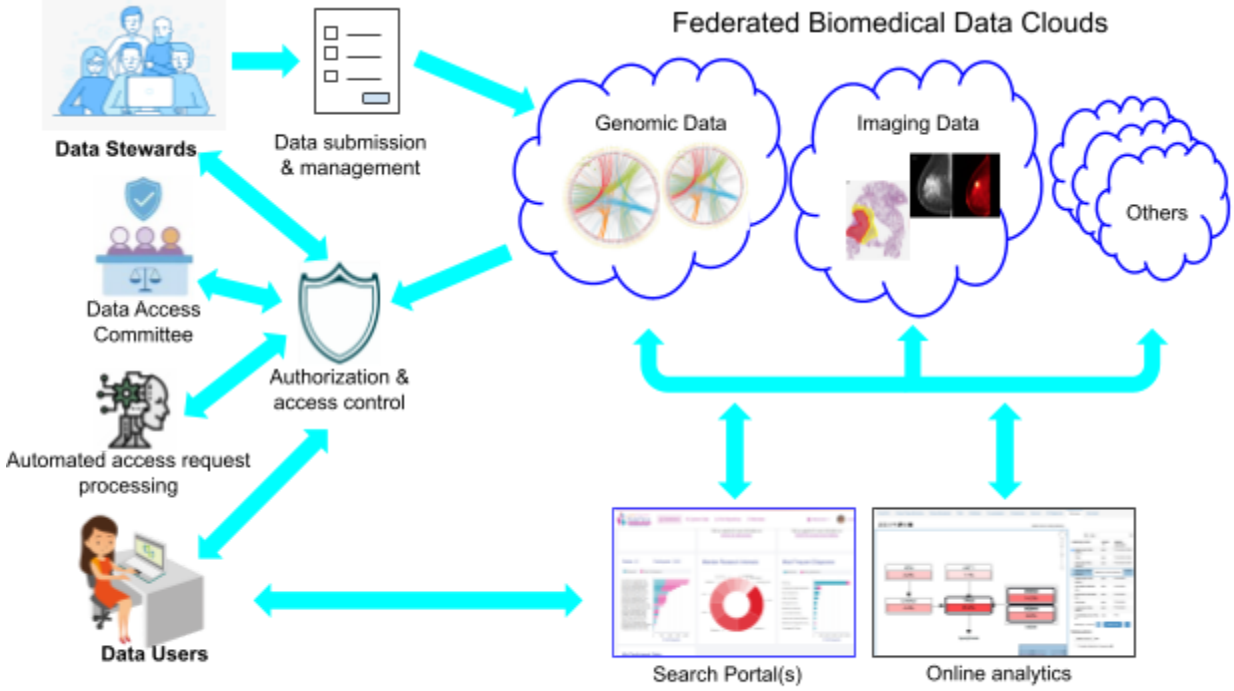

Figure 1: The biomedical data cyberinfrastructure provides the building blocks for securely storing, sharing & analyzing legacy biomedical data sets

Use of published standards and compute cloud technologies

The proposed digital infrastructure will use the biomedical data sharing standards that have been formalized by international organizations such as the Global Alliance for Genomics & Health (GA4GH).[2] Furthermore, to facilitate adherence to FAIR principles, the software will be designed to run on top of commercial and/or academic compute clouds. The advantage of using clouds is that they allow communities to stand up new data repositories easily, and for the repositories to scale in a cost-effective fashion by taking advantage of the virtually limitless capacity of commercial clouds.

By using clouds, repositories built on this system can support multiple data access models including: (1) traditional upload/download; (2) data analytics executed within the same cloud that the data is stored in; (3) data analytics executed within a different cloud using software that transparently pulls the data from one cloud to another in a just-in-time fashion; or (4) a federated model in which different parts of the repository

are located in different clouds, and software moves the data and/or the analytic software among them as needed. Federated models are particularly useful when portions of the data set must reside in particular locations in order to satisfy legal requirements.

Lastly, the high level of security provided by compute clouds, which includes end-to-end data encryption, the option of utilising cloud data centres situated on Canadian soil, and the ability of researchers to analyze data within the same cloud that the data resides in, will allow us to offer a safe harbor for the storage and analysis of Canadian personal health information in compliance with PHIPA[3] and similar patient protection regulations.

Modular submission and federated search

The submission system, search portal and data access control system are the three key software elements of the proposed digital infrastructure. The *submission system* will consist of several modular software elements that facilitate the upload and validation of data objects and associated metadata. Data can be submitted interactively using a web interface, in batch using command line tools and Application Programming Interfaces (APIs), or in combination. Metadata elements that describe the nature of the data object will accompany the object, and both data and metadata will be validated against a *data dictionary* defined by the community sponsoring the repository in question. The infrastructure will provide a default core data dictionary containing a common set of fields and validation rules including the provenance of the object, its access policies, and project information to allow the object to be linked to publications, as well as a set of standard fields for genomic, clinical and imaging data.

When submitted and validated, each data object is issued an immutable globally unique identifier (GUID), which uniquely identifies the object across all repositories. A data management interface will allow the data steward (the individual or group responsible for granting access) and their delegates to track the submission process, edit the metadata, and manage the object's visibility and sharing settings. The submission system's data management interface will optionally allow a data curation step in which trained curators manually review and clean up submissions prior to their acceptance.

The metadata for each visible data object will be indexed for rapid search and retrieval by a network of federated *search portals.* Using the portal of their choice, a researcher will be able to rapidly search across all repositories to find data objects that meet their criteria, generate a manifest of GUIDs that can be passed to analytic software for retrieval, analysis and integration. An example of such a search interface can be found at the International Cancer Genomics Consortium's data portal developed by OICR[4].

Support for multiple data governance and data access models

A key aspect of this concept is its support for multiple data governance models. The communities sponsoring a repository will be able to pick from several governance models including: (1) the data steward grants access to individual researchers; (2) a designated Data Access Committee grants access to individual researchers or research groups; (3) registered access to qualified researchers[5]; and (4) the Data Use Ontology (DUO)[6]. The latter two protocols provide for rapid automated decision making on

whether a researcher is allowed to access a particular data object. The first does so by checking whether a researcher has been certified by her institution as having the training needed to access certain classes of confidential, but low risk data, such as anonymized genome sequences. The second provides a mechanism by which machine-readable descriptions of the acceptable uses of a data object can be matched against machine-readable descriptions of the purposes that the researcher wants to put it to. Both registered access and DUO-based matching are enabled by the GA4GH Passport system, a recently-approved technical standard for federating governance of research data.[7] For identity management, the proposed data access and authorization system will take advantage of several established federated systems, including LinkedIn and ORCID academic author IDs.[8]

From concept to implementation

This proposal is allied with other whitepapers that have been submitted in response to the NDRIO call, including "*Digital research infrastructure to support federated computing on large scale biomedical datasets*" (Brudno *et al*) and "*Canada's Future DRI Ecosystem: AI Research Needs*" (Bodkin *et al*). The concept is supported by industry cloud providers, and leverages work begun by regional and national digital infrastructure initiatives such as the Ontario Health Data Platform, Canada's Digital Supercluster and the Terry Fox Digital Health Discovery Platform. The next steps are to bring together users, data providers, and funding agencies to discuss a suitable set of high-value use cases, and then to mount a proof of principle project to address a high-priority need. Projects under discussion include registries of data related to rare and hereditary diseases in order to facilitate early diagnosis and treatment, and a repository of medical images to support the development of machine learning algorithms for precision medicine. Both of these applications are ripe for commercialization and health care impact. The time is now, and the goal within our reach.

Literature cited

1. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
2. Birney, E., Vamathevan, J. & Goodhand, P. Genomics in healthcare: GA4GH looks to 2022. *Cold Spring Harbor Laboratory* 203554 (2017) doi:10.1101/203554.
3. Beardwood, J. P. & Kerr, J. A. Coming soon to a health sector near you: an advance look at the new Ontario Personal Health Information Protection Act (PHIPA): part II. *Healthc. Q.* **8**, 76–83 (2005).
4. Zhang, J. *et al.* The International Cancer Genome Consortium Data Portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
5. Dyke, S. O. M., Linden, M. & Lappalainen, I. Registered access: authorizing data access. *European Journal of* (2018).
6. Data Use Ontology approved as a GA4GH technical standard. https://www.ga4gh.org/news/data-use-ontology-approved-as-a-ga4gh-technical-standard/.
7. Dyke, S. O. M. Chapter 2 - Genomic data access policy models. in *Responsible Genomic Data Sharing* (eds. Jiang, X. & Tang, H.) 19–32 (Academic Press, 2020).
8. Haak, L. L., Fenner, M., Paglione, L., Pentz, E. & Ratner, H. ORCID: a system to uniquely identify researchers. *Learn. Publ.* **25**, 259–264 (2012).