PIP Digest

CCRA ACRC Canadian Cancer Research Alliance / Alliance canadienne pour la recherche sur le cancer

PIP PPP Patient Involvement in Cancer Research Program / Programme de Participation des Patients à la recherche sur le cancer

CANCER RESEARCH: Big Data and Precision Health

This PIP Digest explains the role of big data in helping to facilitate precision health.

Key Concepts

- Big data
- "Data to knowledge to action" cycle
- Big data challenges
- Key initiatives

Related PIP Digests

- Clinical Trials: Precision Medicine and Clinical Trials
- Cancer Research: Cancer Research Funding

Cancer research has always been data-driven, but two decades of explosive advances in data collection, storage, and analysis have entirely changed the game, allowing researchers to understand much more about both the molecular biology of cancer and about each individual who develops the disease.

The term "big data" reflects not only how much more data researchers now have access to, but also to the ways that massive data sets can be combined to reveal unexpected connections, hidden risk factors, and interventions tailored to subgroups and even to individuals.

On its own, big data is not actually that useful. Putting it to work requires sophisticated analysis that often exceeds human capacity and requires artificial intelligence to extract insights. These insights drive new generations of "precision health," moving from "data to knowledge to action" (as depicted on the following page).

Big data "captures the opportunities and challenges involved with accessing, managing, analyzing, and integrating information within diverse data sets … These data sets currently exceed the abilities of traditional data management approaches."[1]

Big data is characterized by[2]:

- the volume and pace at which it is generated
- multiple data sources and formats, involving different locations, times, intervals, and methodology
- the inclusion of data sources that may be incomplete and inaccurate

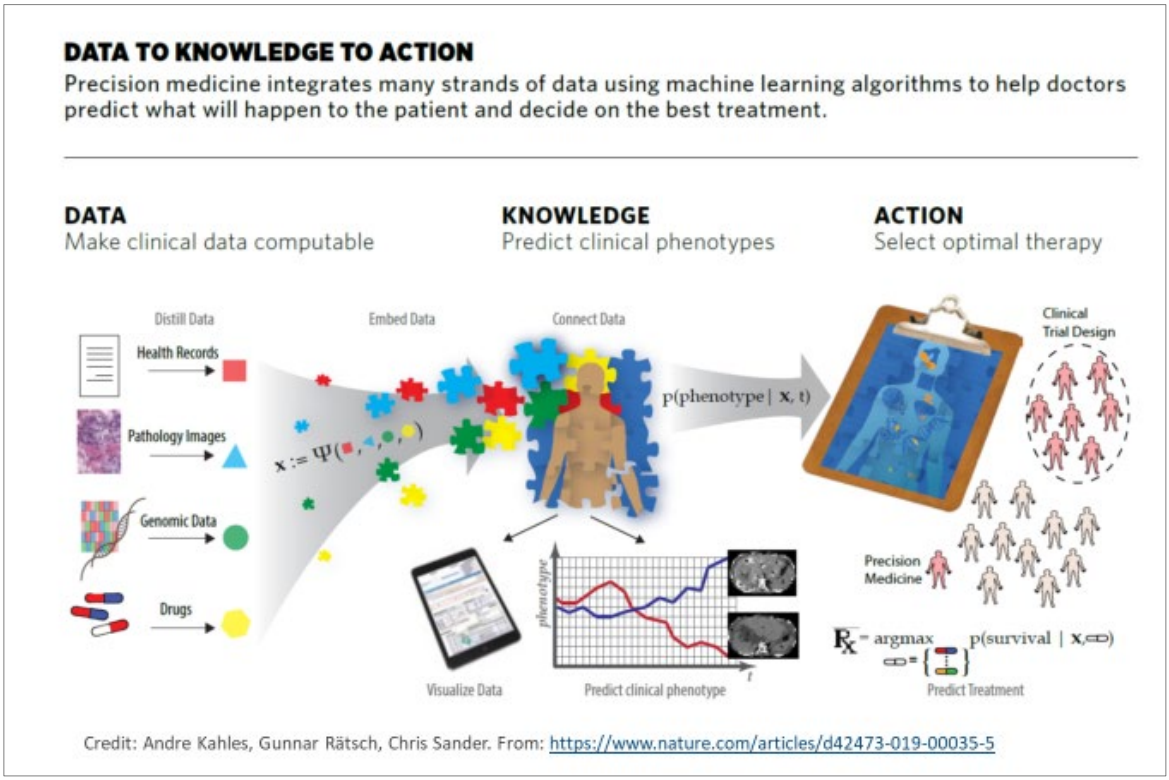
[1]From: https://cancercontrol.cancer.gov/brp/priority-areas/big-data.html.
[2]Willems SM et al. (2019). The potential use of big data in oncology. *Oral Oncology*, 98:8-12.

- the capability to use the data both to see the "big picture" of cancer research, and also to reach down to the granular level of each affected individual

Researchers use computer algorithms to analyze these huge, varied datasets. For example, a researcher may wish to predict the long-term impact of specific chemotherapies on cancer patients. The right algorithm could process millions of electronic health records — each containing hundreds of data points — to identify how specific patient characteristics relate to specific outcomes.

In the health setting, sources of big data may include:

- physicians' offices
- hospitals
- pharmacies
- laboratories/test centres (including genetic and other testing, imaging and pathology results)
- health insurance providers
- wearable monitors, sensors and devices, including smartphones
- social media
- web search information/patterns
- government records (census, vital statistics, disease registries, surveillance)
- research studies, including, cohort studies, clinical trials, and large sequencing studies



Credit: Andre Kahles, Gunnar Rätsch, Chris Sander. From: https://www.nature.com/articles/d42473-019-00035-5

While it can be challenging to extract meaning and insight from such a varied combination of data sources, researchers have better and better tools to do so.

Big data has the potential to:

- detect cancers sooner, allowing for early intervention
- assess unique risk factors for subgroups and individuals
- tailor treatment to each patients' genome, medical history, and other personal factors
- identify unexpected connections between genetic, environmental, socioeconomic, and other factors that affect cancer risk, cancer development and interventions to prevent and reduce cancer

Big Data Challenges

Data Standards

Many existing sources of health data were created and stored by specific institutions and organized without consideration of how they might "speak" to other data sources. Consolidating and harmonizing such datasets remains one of the cancer-research communities' biggest challenges.

Developing more universal data standards requires the expertise of scientists from many disciplines, as well as health care providers, data and computer scientists, software engineers, ethicists, decision-makers, and industry researchers.[3] It also involves logistical challenges like developing standards that can cross jurisdictional boundaries.

Infrastructure capacity

Big data infrastructure involves the tools, people, and systems for collecting, storing, securing, retrieving, sharing, and analyzing data. Aside from the technological challenges, privacy and security, intellectual property, legality, and ethics must all be considered.

Big data requires and drives many technological innovations. Cloud platforms, compression algorithms, data indexing, user interfaces and visualization tools all help to make big data more manageable, accessible, and useful. Open-source resources with strong documentation help ensure these resources are democratized, and that everyone can benefit from the insights they provide.[4]

Stewardship

In Canada, the big data "ecosystem" requires thoughtful approaches by federal and provincial governments to develop unified policies around health and health-related data. They need preserve the security and privacy of individual health data and ensure that new research reduces rather than expands health disparities.[5] This seemingly highly technical type

[3]Hulsen T et al. (2019). From Big Data to Precision Medicine. *Frontiers in Medicine*, 6(34). https://www.frontiersin.org/articles/10.3389/fmed.2019.00034/full

[4]Hinkson IV et al. (2017). A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. *Frontiers in Cell Development and Biology*, 5(83). http://dx.doi.org/10.3389/fcell.2017.00083.

[5]Vayena E et al. (2018). Policy implications of big data in the health sector. *Bulletin of the World Health Organization*, 96(1):66-8. http://dx.doi.org/10.2471/BLT.17.197426

of research, still needs to put people at the centre of policy and practice. Empowerment means giving participants choices and the tools to understand their potential risks and benefits.

Key Initiatives

As data becomes ever-more complex, capacity-building is crucial. **Bioinformatics.ca** offers advanced training workshops on bioinformatics, genomics, proteomics, and technologies related to computational biology. For more information, see https://bioinformatics.ca/.

The **Cancer Genome Collaboratory** is an academic research cloud built by the Ontario Institute for Cancer Research (OICR). It is home to the data holdings of the International Cancer Genome Consortium. Using the Collaboratory's facilities, researchers can run complex data mining and analysis operations across a large repository of cancer genome sequences and their associated donor clinical information. For more information see https://www.cancercollaboratory.org/.

Compute Canada provides essential advanced research computing services and infrastructure to Canadian researchers and their collaborators in academic and industrial sectors. In collaboration with regional organizations, they help researchers use big data to accelerate research and innovation. For more information, see https://www.computecanada.ca/.

The **Global Alliance for Genomics and Health** (GA4GH) is an international, non-profit alliance working to accelerate research and medicine to advance human health. Bringing together more than 500 organizations in healthcare, research, patient advocacy, life science, and information technology from more than 90 countries, the GA4GH creates frameworks and standards to enable responsible, voluntary, and secure sharing of genomic and health-related data. For more information, see https://www.ga4gh.org/.

ICES is an independent, non-profit corporation that applies health informatics to health-services and population-wide health outcomes research in Ontario. Using data collected through the province's public healthcare system, their data repository consists of patient-level, coded and linkable health records. Through partnerships, the repository also securely links data from a variety of health surveys and registries. By linking the different data sets together using anonymous numeric unique identifiers, ICES scientists track different aspects of provincewide health service use and patient outcomes over time. For more information, see https://www.ices.on.ca/. [Note that other provinces also have linked data and processes for accessing population-level, health-related data.]

Check out these short videos for more information on big data:

- Dalhousie FCS. *Big Data Institute*. (YouTube) March 30, 2016 [4:16 minutes] https://www.youtube.com/watch?time_continue=96&v=ibtR1dHkebM
- École Polytechnique. *How data science can help doctors fight cancer.* (YouTube) Jun 29, 2018 [2:18 minutes] https://www.youtube.com/watch?v=TRkKnnYgfks
- GA4GH. *GA4GH: The Global Standards Organization for Genomics*. (YouTube) 2019 [2:09 minutes] https://www.youtube.com/channel/UCmCg1AcAY_qHXfOhePAFAeQ
- HuffPost. *How Big Data Could Transform the Health Care Industry.* (YouTube) February 2, 2017 [3:48 minutes] https://www.youtube.com/watch?v=_mXrZEIpNMw
- ICES Ontario. *Institute for Clinical Evaluative Sciences (ICES): The power of data to improve health*. (YouTube) October 5, 2016 [2:39 minutes] https://www.youtube.com/watch?v=5pZhRSM1cyk&feature=youtu.be
- UHNToronto. *Lillian Siu talks 'big data' in cancer prevention and care*. (YouTube) September 1, 2016 [2:54 minutes] https://www.youtube.com/watch?v=1JhSGPQE3_0

Last revised: 2020-03-26