Progress Towards Cancer Data Ecosystems

Robert L. Grossman
University of Chicago

Accelerating Research Through Data Sharing

With the new instruments producing ever larger amounts of data relevant to biology, medicine and healthcare (1) and with the maturing of data science (2), there are increasing opportunities to use data science to make discoveries. Sharing cancer data can provide the data required to make basic discoveries about the mechanisms of cancer (3), accelerate the discovery of biomarkers (4), inform the repurposing of drugs (5), and aggregate data to provide the evidence required to discover genomic alterations associated with rare subtypes of cancer or other rare diseases (6). An overview of the importance of sharing molecular data in cancer can be found in (7). A succinct target for data sharing in cancer was expressed by former Vice President Biden with the phrase "The goal […is] to accelerate progress in cancer research and treatment enough to do in 5 years what would have taken 10 [years]" (8).

A fundamental tension in biomedical research is the trade-off between: i) ethical, legal and regulatory requirements to obtain consent from patients and to protect patients and their data; and ii) the right of patients to benefit from research (9). See Figure 1. There are always potential risks when different datasets and large datasets are brought together (10, 11). For example, through what are sometimes called side channel attacks (12), it is becoming increasingly easy to uncover the likely identities of *some* individuals when large deidentified datasets are made available. An example in (13) shows how genotype data, general geo-location information (say associated with the hospital that provided the biospecimen), and a local white pages can be used to create a short list of individuals that may have potentially contributed data to the 1000 Genomes Project.

One of the ten recommendations of the Cancer Moonshot Blue Ribbon Panel (14) was to build a national cancer data ecosystem: "…that will enable all participants across the cancer research and care continuum to contribute, access, combine and analyze diverse data that will enable new discoveries and lead to lowering the burden of cancer in our country (14)." The Blue Ribbon Panel report lists as one of the goals of a cancer data ecosystem is to: "allow both public and private information resources to be readily discovered and connected through the use of a common information architecture (14)."

In this review, we address two questions related to information architectures for supporting a cancer data ecosystem:

1. What are some of the systems that can be used for sharing cancer genomic and clinical data with security and compliance in order to gain access to a critical mass of data that can shorten the time required to make discoveries that benefit cancer patients?

2. How can we build an ecosystem of interoperating systems for sharing data and applications that access them so that research scientists can most effectively use the data that is being shared?

Data Clouds and Data Commons

Data commons usually rely on cloud computing infrastructure, which is based upon the virtualization (15) of computing, storage, networking and other services required for computing. Virtualization provides the required computing infrastructure using software simulation rather than physical hardware. This provides two main advantages: first, different simulated hardware configurations can use the same underlying physical hardware; second, the underlying physical hardware can be used more efficiently. With this approach, storage and computing resources can be provisioned elastically as needed by users, even at large scale (16). Cloud computing infrastructure is available from open source software, such as OpenStack (17), that can be run by an organization locally or on products from third party cloud service providers, such as Amazon's AWS, Google's GCP, or Microsoft's Azure.

For the purposes here, we can define a *data commons* as a system that co-locates data with cloud computing infrastructure and commonly used software services, tools and applications for managing, analyzing and sharing data to create an interoperable resource for the research community (18).

Some of the most important services for working with cancer data include: 1) authentication and authorization services, 2) digital ID services; 3) metadata services; and 4) services for launching virtual machines, containers, notebooks or similar infrastructure for computing over data.

Additional services that are sometimes supported include: 5) using a common data dictionary and integrating data with respect to a common data model; 6) processing the integrated data with respect to a common set of bioinformatics pipelines; 7) exposing the integrated and harmonized data using an API. Notice that services 1) – 3) when exposed via an API, as is standard, is enough to make data *findable* and *accessible* as usually defined (19), while 5) is enough to make data *interoperable* (19).

The term *data commons* is beginning to be used to a refer to a system with capabilities 1) – 7) and the term *data cloud* to a refer to system when properties 1) – 4) are supported.

One of the important use of data clouds is to support the large-scale processing of NGS data. Researchers can define their own bioinformatics workflows, work with members of an analysis

working group (AWG) to develop workflows for the working group, or use workflows developed and optimized by third party vendors.

Data clouds that support user-defined workflows to support individual researchers or analysis working groups include: the University of Chicago's Bionimbus Protected Data Cloud (20), OICR's Cancer Collaboratory (21), the Galaxy Cloud (22, 23), Seven Bridges Cancer Genomics Cloud (24), Broad's FireCloud (25), and the Institute for Systems Biology Cancer Genomics Cloud (26). The latter three were part of the NCI Cloud Pilots (now Cloud Resources) (27). A good analysis of using multiple clouds to support an analysis working group is in (28).

Data clouds that support running vendor supplied bioinformatics pipelines (sometimes called *NGS as a service*) include the DNAnexus platform (29, 30) and Globus Genomics (31).

The NCI Genomic Data Commons and the BloodPAC Data Commons (32) are examples of a data commons satisfying requirements 1) - 7) (33, 34). Data commons usually have portals for exploring data, downloading data, as well as submitting data, and APIs so that third party software services, applications, notebooks (35) and cloud-based collaborative workspaces can access data managed by the commons.

Pediatric Data Commons

Due to the much small number of patient data available within the pediatric cancer research community, data sharing is essential for many discoveries (36). As an example, the International Neuroblastoma Risk Group (INRG) collects clinical data and genotype data about neuroblastoma tumors from patients from around the world to create a database and associated data commons (37). Given the relatively small number of neuroblastoma tumors each year, without such an effort there would not be enough data to reliably classify neuroblastoma tumors (38).

Other examples of data commons supporting the pediatric cancer community include Cavatica (www.cavatica.org) and the Kids First Data Resource (d3b.center/kidsfirst/).

Standards

Standards for data commons and data ecosystems are immature and still emerging. Efforts worth mentioning include the Global Alliance for Genomics and Health (GA4GH) (39), which is establishing standards for sharing clinical and genomic data within frameworks for data governance, security, privacy and patient benefits. The FORCE11 FAIR principles (19) describe guidelines so that data can be findable, accessible, interoperable and reusable (FAIR).

A challenge with efforts to establish standards for data commons is that the standards can get so far ahead of any scalable implementations that they become irrelevant. In response to this, three large scale projects for sharing clinical and genomic data began to collaborate on APIs,

while competing on building platforms. The three groups are some of the developers behind the i) NCI Genomic Data Commons and NCI Cancer Research Data Commons; ii) the NIH All of Us Research Program (https://allofus.nih.gov/); and iii) the Human Cell Atlas (HCA) (40). The collaboration is called the Commons Alliance or Data Biosphere. It is important to note that it is a commitment of a group of the core software developers supporting these projects that are working together to collaborate on standardized APIs, and not, at this time, a commitment of the organizations themselves. Note that a similar approach was taken by the NCI Cloud Pilots ("collaborate on standardized APIs and compete on implementations") (27).

Cancer Data Ecosystem – Initial Steps

Currently, one way that cancer data ecosystems are emerging *is the ability for several data commons to interoperate and to share a common ecosystem of applications by each using a core set of common services*. For example, the NCI GDC (33), the BloodPAC Data Commons for liquid biopsies (32), and the Data Catalog for the Kids First Data Resource, plus several additional data commons, can all interoperate because they are built with the same open source software (Gen3) and are built around the following core services that are all accessible through open APIs:

1. Authentication and authorization services
2. Digital ID services
3. Metadata services

In addition, they all use the NCI Thesaurus (41), and are based in part on the GDC Data Model and use the GDC API (34). With the digital ID services, data can be in any private or public cloud, located from its metadata 3) using any convenient searching and indexing services, and accessed by a client through an API, as long as the user is authenticated and authorized to access the data using 1).

More information about the Gen3 software stack for building data commons and data ecosystems can be found in (42).

The NCI Cancer Data Ecosystem

Perhaps the most mature cancer data ecosystem has been developed by the National Cancer Institute and currently consists of the NCI Genomic Data Commons and the NCI Cloud Resources (27). It is called the NCI Cancer Research Data Commons (NCRDC). The Cloud Resources (43) in the NCRDC use AWS and the GCP so that users can compute over data from the GDC and other resources and contain different applications for analyzing data. Importantly, the Cloud Resources also support workspaces. Additional commons are being added to the NCRDC including a proteomics and imaging commons, with additional commons and resources in planning. The NCRDC Framework Services can be used so that the different data commons and other NCRDC resources can share authentication, authorization, ID and metadata services.

More specifically, the NCRDC data ecosystem is based on the following framework services:

1. Authentication (AuthN) and Authorization (AuthZ) services
2. Metadata validation using the NCI Thesaurus (41) and related services
3. An extensible data model based upon the GDC data model (34)
4. APIs for containers and workflows based upon the GA4GH WES and TES standards
5. Services to set up and access collaborative workspaces.

In particular, this enables FAIR resources to be built and for the data commons and resources to interoperate.

Currently, public clouds typically charge egress charges for moving data outside of the cloud boundary. Network peering is the idea that two internet service providers exchange network traffic at no charge, which is in contrast to the charges imposed when an internet service provider charges its customers for transmitting network traffic (44). Data peering between the large internet service providers was an important driver for the growth of the internet. An important goal for building a data ecosystem is to support *data peering* between data commons so that research data can be accessed from one data commons at no cost, even when it is stored in another data commons or cloud (18). Since public clouds make money when their users compute over data, supporting some level of data peering should be quite feasible.

There is no agreed to definition of a data ecosystem at this time, but, at the minimum, a *data ecosystem* (as opposed to a data commons) should support:

1. Authentication and Authorization services so that a community of researchers can access an ecosystem of data and applications with a common (research) identity.
2. The ability for multiple data commons to interoperate, preferably through data peering.
3. Shared data models to simplify the ability for third party applications to access data from multiple data commons.
4. A collection of applications that are powered by APIs that are FAIR compliant.

A properly designed and operated data commons is not limited by the type of data, the amount of data, or the type of applications that can be supported, but usually by the priorities, interests and funds of the sponsors and operators of the data commons. For this reason, one of the most important roles of a data ecosystem is to support multiple data commons that can interoperable and a rich variety of third party applications over them.

Architectures for Data Ecosystems to Support Cancer Research

There is no consensus system architecture at this time for the key components in a data ecosystem to support cancer research. Below, we discuss four options that are currently being used.

Commons with an open API. With this approach, a data commons exposes an open API, which can be used by large scale cloud resources. This was the approach that emerged during 2014-2016 with the NCI Genomic Data Commons (GDC) (33) and NCI cancer clouds developed by Broad (25), Seven Bridges Genomics (24), and the Institute for Systems Biology (26). This architecture is enabled by the GDC API (34), which is used both by the GDC itself for all its applications, by the three cancer clouds, and by an increasing number of third party applications.

Walled Garden. Another approach is to develop a data ecosystem within a proprietary system. Examples include DNAnexus (29) and Seven Bridges Genomics (24). These types of systems may expose APIs for users and may be part of a broader ecosystem, but are not themselves open source.

Narrow middle design. An important design philosophy when designing complex systems is sometimes called the *end-to-end principle*, "[which argues] for moving function upward in a layered system, closer to the application that uses the function (45)." The underlying architecture for the internet was based in part on this principle, with the end result that the protocols such as tcp and http in the middle were relatively simple and changed slowly, in contrast to the internet applications at the "ends" which were much more complicated and evolved rapidly (45). In our context of data ecosystems for cancer, we might term this the *narrow middle design*, in which the minimal cores services necessary to support a data ecosystem are kept few, small and simple, with innovation and rich applications at the ends: for getting data into the ecosystem (at the "input end") and for exploring, integrating, harmonizing, and analyzing data in the ecosystem (at the "output" end). Notice that the data commons framework services in the National Cancer Research Data Commons is an example of this approach. See Figure 3.

Modular components and services. Another approach being explored for building data commons and ecosystem is to support enough core modular services and components that are FAIR compliant (19) that a complete commons or data ecosystem can be built. For example, the NIH Data Commons Pilot Phase Consortium (DCPPC) is taking this approach and developing a system by integrating a suite of components and services, including persistent identifier services, metadata services, data indexing services, data search services, services to support workspace, and services for working with complex phenotype and genotype data (commonfund.nih.gov/commons).

Once multiple systems have been developed, it is common to federate them (46) in order to create a large ecosystem. There are multiple ways that this can be done (46) and Table 1 describes some of them.

Design	Example
Single system with open APIs supporting third party applications	NCI Genomic Data Commons with third party applications, notebooks and workspaces (34)
Federation of multiple independent interoperating systems with APIs supporting third party applications	The Matchmaker Exchange Platform for Rare Disease Gene Discovery (6)
Federation of multiple independent data commons with common services and common core data models, with open APIs supporting third party applications	BloodPAC Data Commons (32) and other Gen3 data commons (42) and applications that can access data from them
Federation of multiple independent interoperating systems over a set of common services and common core data models, with open APIs supporting third party applications	NCI Cancer Research Data Commons, including GDC, Cloud Resources, Proteomics Databases (27)
Federation of multiple independent systems connected via APIs	Kids First Data Resource (d3b.center/kidsfirst)

Table 1: Some of the different federation architectures that have been used for building data ecosystems for cancer.

Cancer Data Ecosystems – Next Steps

Moving from the current state of cancer data ecosystems requires making three transitions:

First, the various current systems for sharing cancer data, including the NCI GDC, NCI Cloud Resources, NCBI's ClinVar (47), AACR's Project GENIE (48) , ASCO's CancerLinQ (49), etc. must themselves interoperate. See Figure 2. These systems are organized into different "lanes," with different regulatory environments and different incentives for sharing and operated by different entities with different sustainability models, which makes interoperating more challenging

Second, patient portals and related infrastructure supporting patient partnered research must be integrated into all aspects of the ecosystem. This technology is still maturing, and which systems will emerge as the dominant ones supported by large groups of patients is still not at all clear.

Third, changes must be made to make it easier to get data into the ecosystem and to curate the data that is part of it. This effort requires both better tools and a larger better trained workforce of bioinformaticans and data scientists.

Fourth, sustainability is both a major challenge and a major reason that current systems do not interoperate. Today's systems often limit access to dues paying members or sign commercial agreements that limit how data can be shared. Funders, including federal agencies and foundations, must build data sharing and support for data sharing ecosystems into how they support research, or we risk not having the critical mass of data available that is needed to improve patient outcomes. Payers must also insist on data sharing so that the required evidence is available to inform standard of care instead of relying on black box systems that do not share the underlying evidence that is used for decision support.

References

1. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big data: astronomical or genomical? PLoS Biol. 2015;13(7):e1002195.
2. Donoho D, editor. 50 years of Data Science. Based on a Presentation at the Tukey Centennial Workshop; 2015: NJ Princeton.
3. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113.
4. Khleif SN, Doroshow JH, Hait WN. AACR-FDA-NCI Cancer Biomarkers Collaborative consensus report: advancing the use of biomarkers in cancer drug development. Clinical Cancer Research. 2010;16(13):3299-318.
5. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. Briefings in bioinformatics. 2015;17(1):2-12.
6. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: a platform for rare disease gene discovery. Human mutation. 2015;36(10):915-21.
7. Siu LL, Lawler M, Haussler D, Knoppers BM, Lewin J, Vis DJ, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. Nature medicine. 2016;22(5):464.
8. Ferber D. Biden blasts science denialists, calls to dramatically speed up fight against cancer. Science Magazine. 2018 February 19, 2018.
9. Knoppers BM. Framework for responsible sharing of genomic and health-related data. The HUGO Journal. 2014 2014/10/17;8(1):3.
10. Karp DR, Carlin S, Cook-Deegan R, Ford DE, Geller G, Glass DN, et al. Ethical and practical issues associated with aggregating databases. PLoS medicine. 2008;5(9):e190.
11. Caulfield T, McGuire AL, Cho M, Buchanan JA, Burgess MM, Danilczyk U, et al. Research ethics recommendations for whole-genome research: consensus statement. PLoS Biol. 2008;6(3):e73.
12. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nature Reviews Genetics. 2014;15(6):409-21.
13. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339(6117):321-4.
14. Panel BR. Cancer Moonshot Blue Ribbon Panel Report. 2016 [cited 2018]; Available from: https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative.

15. Scarfone K, Souppaya M, Hoffman P. Guide to Security for Full Virtualization Technologies, National Institute of Standards and Technology (NIST), January 2011.
16. Mell P, Grance T. The NIST Definition of Cloud Computing (Draft): Recommendations of the National Institute of Standards and Technology. National Institute of Standards and Technology, 2011.
17. Pepple K. Deploying openstack: " O'Reilly Media, Inc."; 2011.
18. Grossman RL, Heath A, Murphy M, Patterson M, Wells W. A Case for Data Commons: Toward Data Science as a Service. Comput Sci Eng. 2016 Sep-Oct;18(5):10-20.
19. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
20. Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, et al. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. J Am Med Inform Assoc. 2014 Nov-Dec;21(6):969-75.
21. Yung CK, Mihaiescu GL, Tiernay B, Zhang J, Gerthoffert F, Yang A, et al. The Cancer Genome Collaboratory. AACR; 2017.
22. Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J, editors. Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics; 2010: BioMed Central.
23. Afgan E, Baker D, Coraor N, Goto H, Paul IM, Makova KD, et al. Harnessing cloud computing with Galaxy Cloud. Nat Biotechnol. 2011;29(11):972.
24. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, et al. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research. Cancer Res. 2017 Nov 1;77(21):e3-e6.
25. Birger C, Hanna M, Salinas E, Neff J, Saksena G, Livitz D, et al. FireCloud, a scalable cloud-based platform for collaborative genome analysis: Strategies for reducing and controlling costs. bioRxiv. 2017:209494.
26. Reynolds SM, Miller M, Lee P, Leinonen K, Paquette SM, Rodebaugh Z, et al. The ISB Cancer Genomics Cloud: A Flexible Cloud-Based Platform for Cancer Genomics Research. Cancer Res. 2017 Nov 1;77(21):e7-e10.
27. Hinkson IV, Davidsen TM, Klemm JD, Kerlavage AR, Kibbe WA. A Comprehensive Infrastructure for Big Data in Cancer Research: Accelerating Cancer Research and Precision Medicine. Front Cell Dev Biol. 2017;5:83.
28. Yung CK, O'Connor BD, Yakneen S, Zhang J, Ellrott K, Kleinheinz K, et al. Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments. bioRxiv. 2017.
29. Reid JG, Carroll A, Veeraraghavan N, Dahdouli M, Sundquist A, English A, et al. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics. 2014;15(1):30.
30. Shringarpure SS, Carroll A, Francisco M, Bustamante CD. Inexpensive and highly reproducible cloud-based variant calling of 2,535 human genomes. PloS one. 2015;10(6):e0129277.
31. Madduri RK, Sulakhe D, Lacinski L, Liu B, Rodriguez A, Chard K, et al. Experiences building Globus Genomics: a next-generation sequencing analysis service using Galaxy, Globus,

and Amazon Web Services. Concurrency and Computation: Practice and Experience. 2014;26(13):2266-79.
32. Grossman RL, Abel B, Angiuoli S, Barrett JC, Bassett D, Bramlett K, et al. Collaborating to Compete: Blood Profiling Atlas in Cancer (BloodPAC) Consortium. Clin Pharmacol Ther. 2017 May;101(5):589-92.
33. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med. 2016 Sep 22;375(12):1109-12.
34. Wilson S, Fitzsimons M, Ferguson M, Heath A, Jensen M, Miller J, et al. Developing Cancer Informatics Applications and Tools Using the NCI Genomic Data Commons API. Cancer Res. 2017 Nov 1;77(21):e15-e8.
35. Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, et al., editors. Jupyter Notebooks-a publishing format for reproducible computational workflows. ELPUB; 2016.
36. Volchenboum SL, Cox SM, Heath A, Resnick A, Cohn SL, Grossman R, editors. Data Commons to Support Pediatric Cancer Research. American Society of Clinical Oncology educational book American Society of Clinical Oncology Meeting; 2017.
37. Volchenboum SL, Cox SM, Heath A, Resnick A, Cohn SL, Grossman R. Data Commons to Support Pediatric Cancer Research. Am Soc Clin Oncol Educ Book. 2017;37:746-52.
38. Pinto NR, Applebaum MA, Volchenboum SL, Matthay KK, London WB, Ambros PF, et al. Advances in Risk Classification and Treatment Strategies for Neuroblastoma. J Clin Oncol. 2015 Sep 20;33(27):3008-17.
39. Lawler M, Siu LL, Rehm HL, Chanock SJ, Alterovitz G, Burn J, et al. All the World's a Stage: Facilitating Discovery Science and Improved Cancer Care through the Global Alliance for Genomics and Health. Cancer discovery. 2015 Nov;5(11):1133-6.
40. Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. The Human Cell Atlas: from vision to reality. Nature. 2017 Oct 18;550(7677):451-3.
41. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics. 2007;40(1):30-43.
42. Gen3 Project Team. Gen3 Documentation. 2018 [cited 2018 March 20, 2018]; Available from: https://uc-cdis.github.io/gen3-user-doc/.
43. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, et al. The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. Cancer Research. 2017;77(21):e3-e6.
44. Huston G, editor. Interconnection, peering, and settlements. proc INET; 1999.
45. Saltzer JH, Reed DP, Clark DD. End-to-end arguments in system design. ACM Transactions on Computer Systems (TOCS). 1984;2(4):277-88.
46. Busse S, Kutsche R-D, Leser U, Weber H. Federated information systems: Concepts, terminology and architectures. Forschungsberichte des Fachbereichs Informatik. 1999;99(9):1-38.
47. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2013;42(D1):D980-D5.

48. Consortium APG. AACR project GENIE: powering precision medicine through an international consortium. Cancer discovery. 2017;7(8):818-31.
49. Schilsky RL, Michels DL, Kearbey AH, Yu PP, Hudis CA. Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ. J Clin Oncol. 2014;32(22):2373-9.

Figures

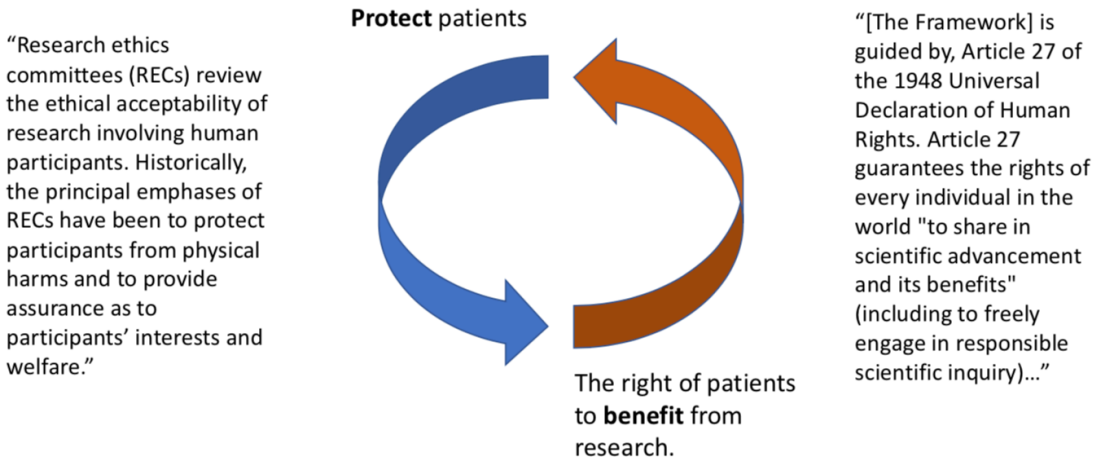

Figure 1. Both quotes in the figure are from (9). One of the challenges with building data ecosystems for cancer is balancing the ethical requirements to protect patients and their data from the right that all individuals have to benefit from scientific advances (9). Today, data ecosystems integrate cancer data that can accelerate research and improve outcomes for cancer patients, but also can lead to potential exposures of data. Balancing these two requires is one of the challenges of building a data ecosystem for cancer.

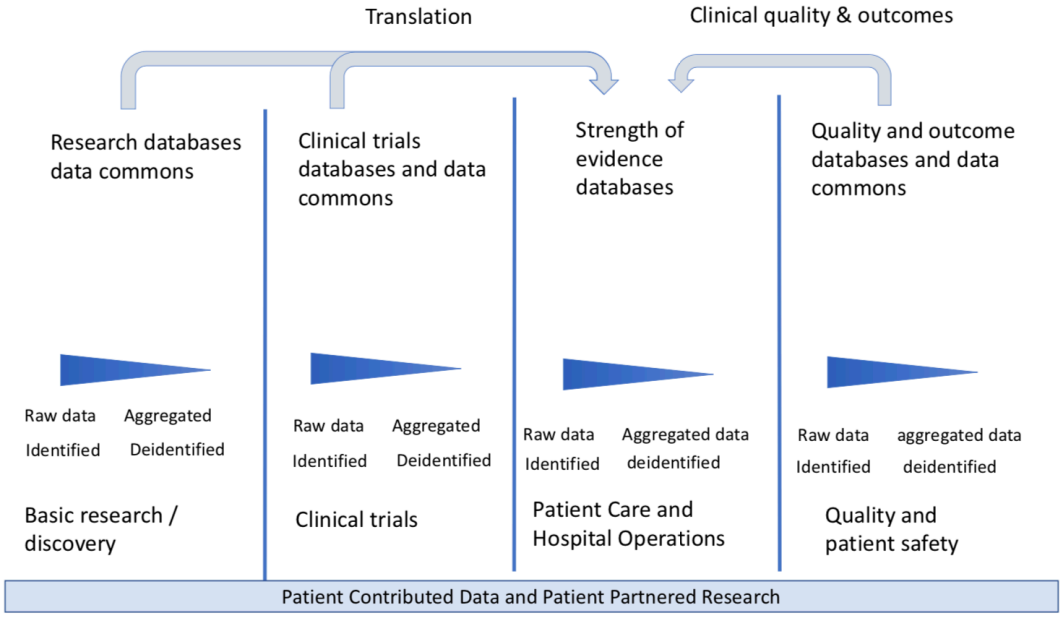

Figure 2. The cancer data ecosystem includes data from: 1) research, 2) clinical trials, 3) patient care and hospital operations, and 4) data quality and outcomes research. An important new source of data will be patient contributed data.

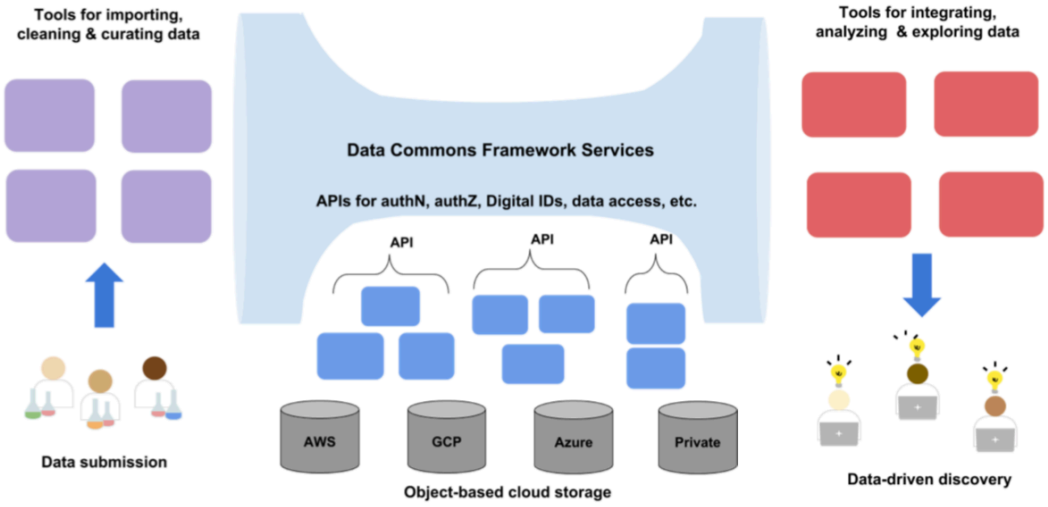

Figure 3. This figure illustrates the "narrow middle" architecture for data ecosystems that uses a core set of framework services that can support a rich collection of rapidly evolving applications and services for ingesting and curating data (at one end) and a rich collection of rapidly evolving applications and services for exploring and analyzing data (at the other end)