Case Studies for Overcoming Challenges in Using Big Data in Cancer

Shawn M. Sweeney[1], Hisham K. Hamadeh[2], Natalie Abrams[3], Stacey J. Adam[4], Sara Brenner[5], Dana E. Connors[4], Gerard J. Davis[6], Louis D. Fiore[7], Susan H. Gawel[6], Robert L. Grossman[8], Sean E. Hanlon[9], Karl Hsu[10], Gary J. Kelloff[11], Ilan R. Kirsch[12], Bill Louv[13], Deven McGraw[14], Frank Meng[15], Daniel Milgram[16], Robert S. Miller[17], Emily Morgan[4], Lata Mukundan[16], Thomas O'Brien[18], Paul Robbins[18], Eric H. Rubin[19], Wendy S. Rubinstein[5], Liz Salmi[20], Teilo H. Schaller[13], George Shi[6], Caroline C. Sigman[15], and Sudhir Srivastava[21]

ABSTRACT

The analysis of big healthcare data has enormous potential as a tool for advancing oncology drug development and patient treatment, particularly in the context of precision medicine. However, there are challenges in organizing, sharing, integrating, and making these data readily accessible to the research community. This review presents five case studies illustrating various successful approaches to addressing such challenges. These efforts are CancerLinQ, the American Association for Cancer Research Project GENIE, Project Data Sphere, the National Cancer Institute Genomic Data Commons, and the Veterans Health Administration Clinical Data Initiative. Critical factors in the development of these systems include attention to the use of robust pipelines for data aggregation, common data models, data deidentification to enable multiple uses, integration of data collection into physician workflows, terminology standardization and attention to interoperability, extensive quality assurance and quality control activity, incorporation of multiple data types, and understanding how data resources can be best applied. By describing some of the emerging resources, we hope to inspire consideration of the secondary use of such data at the earliest possible step to ensure the proper sharing of data in order to generate insights that advance the understanding and the treatment of cancer.

Introduction

Vast amounts of biological and clinical data are being created to provide the research material needed to develop more effective cancer treatments and patient management. In our first paper (1), we described current and evolving principles for managing these data successfully. They need to be organized, shared, integrated, and made readily accessible (2, 3). Our first report highlighted the scope of the challenges associated with each of those steps. It offered an array of existing efforts and opinions aimed at mitigating the respective roadblocks and pain points.

This paper focuses on illustrating the successful implementation and challenges of these efforts through cancer-specific use cases from several major data repositories. These select oncology case studies provide various approaches to overcoming the aforementioned data access, quality, and analytic challenges. Each example starts with a description of the effort's purpose, content, and progress (**Table 1**) and finishes with a discussion of challenges specific to the case studies and lessons learned, summarized in **Table 2**.

Integration of multiple data types and access to analytical tools are the critical capabilities of the resources described in this paper. For example, the Genomic Data Commons (GDC) is a component of the NCI Cancer Research Data Commons (CRDC; https://datascience.cancer.gov/data-commons), which includes multiple sets of curated clinical genomics data, as well as imaging, proteomics, and associated metadata and will soon incorporate digital pathology and multispectral data from the Human Tumor Atlas Network (HTAN; https://humantumoratlas.org).

CancerLinQ

CancerLinQ (https://www.cancerlinq.org/) is a health technology platform developed and implemented by CancerLinQ LLC,

[1]American Association for Cancer Research, Philadelphia, Pennsylvania. [2]Genmab, Princeton, New Jersey. [3]Division of Cancer Prevention, Early Detection Research Network, National Cancer Institute, Rockville, Maryland. [4]Foundation for the National Institutes of Health, Bethesda, Maryland. [5]Office of In Vitro Diagnostics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland. [6]Abbott Diagnostics Division, Abbott Laboratories, Lake Forest, Illinois. [7]Boston University School of Medicine, Boston and New England Department of Veterans Affairs, Bedford, Massachusetts. [8]Center for Translational Data Science, The University of Chicago, Chicago, Illinois. [9]Center for Strategic Scientific Initiatives, National Cancer Institute, Bethesda, Maryland. [10]Sanofi, Bridgewater, New Jersey. [11]Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland. [12]Adaptive Biotechnologies, Seattle, Washington. [13]Project Data Sphere, Morrisville, North Carolina. [14]Ciitizen Platform at Invitae, San Francisco, California. [15]Boston University and Veterans Administration Boston Healthcare System, Boston, Massachusetts. [16]CCS Associates, San Jose, California. [17]CancerLinQ, American Society of Clinical Oncology, Alexandria, Virginia. [18]Pfizer, Brooklyn, New York. [19]Merck, New York, New York. [20]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts. [21]Cancer Biomarkers Research Group, Division of Cancer Prevention, National Cancer Institute, Rockville, Maryland.

Current address for T. Schaller: Kriya Therapeutics, Redwood, California.

Corresponding Authors: Shawn M. Sweeney, American Association for Cancer Research, 615 Chestnut St., Floor 17, Philadelphia, PA 19106. Phone: 215-440-9300; E-mail: shawn.sweeney@aacr.org; and Hisham K. Hamadeh, Genmab US Inc., 777 Scudders Mill Road, Building 2, 4th Floor, Plainsboro, NJ 08536. Phone: 609-455-7501; E-mail: hha@genmab.com

Cancer Res 2023;83:1183–90

doi: 10.1158/0008-5472.CAN-22-1277

This open access article is distributed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.

©2023 The Authors; Published by the American Association for Cancer Research

Downloaded from http://aacrjournals.org/cancerres/article-pdf/83/8/1183/3320604/1183.pdf by guest on 06 August 2026

Table 1. Overview of case studies.

Name of initiative (URL) Lead institution(s) Brief description/mission	Type of data	Size	Database access	CDM(if applicable)	Utility	Ref(s)
CancerLinQ (www.cancerlinq.org) Subsidiary of ASCO Technology platform that collects and analyzes data from patient encounters to improve cancer care quality and advance discovery	Clinical data from EHRs, cancer registry files, and other sources	100+ cancer centers and oncology practices in the USA that have signed on to participate, as of December 2021, representing more than 2 million cancer patients.	Application required for access to CancerLinQ Discovery deidentified research data sets	Quality Data Model	Oncologists in member practices use CancerLinQ dashboards for view of oncology-specific quality measures and to obtain data on treatment specifics for rare cancers, etc. Also, anyone in the cancer community is offered access to distinct sets of aggregated, deidentified patient data with the purpose of investigating specific research questions through an application process.	(4, 5)
AACR Project GENIE Initiative International data-sharing consortium including multiple institutions worldwide (e.g., Dana-Farber Cancer Institute, Institut Gustave Roussy, Netherlands Cancer Institute) Regulatory-grade registry aggregating and linking clinical-grade cancer genomic data with clinical outcomes from tens of thousands of cancer patients treated at participating institutions	Multiple data types. Integrated genomic and clinical data sets.	The 10.0-public release (July 2021) contains 120,953 deidentified genomic records from 111,222 individuals and represents 109 major cancer types and 742 cancer subtypes.	Data accessed directly via cBioPortal for Cancer Genomics or downloaded from the Synapse platform. Users need to create account for either site and agree to terms of access.	Custom, patient-centric CDM using existing standards (e.g., North American Association of Central Cancer Registries, American Joint Committee on Cancer, etc.). Cancer-specific variations made for individual projects/cancer types.	"One Registry, Many Uses"; powers clinical and translational research; validates biomarkers; drug repositioning/repurposing; adds new mutations to existing drug labels; identifies new drug targets; and could provide the evidence base necessary to support reimbursement for NGS-based testing by payers. Data from the registry were recently used to support the regulatory filing for sotorasib in NSCLC.	(6, 8)
Project Data Sphere Independent, not-for-profit initiative of the CEO Roundtable on Cancer; hosting and analytical tools provided by SAS Institute Open-access digital library laboratory that provides a secure platform for researchers to access and analyze deidentified patient-level clinical data, with a focus on late-stage clinical trials	Patient health data. Raw, deidentified clinical and imaging data from late-phase oncology clinical trials.	200+ clinical trial data sets representing 240,000+ patients	One-time user registration gives open access to all data and tools.	Some in CDISC Data Tabulation Model format	Allows researchers to access and combine multiple clinical data sets for analysis	(9–13)

(Continued on the following page)

Downloaded from http://aacrjournals.org/cancerres/article-pdf/83/8/1183/3320604/1183.pdf by guest on 06 August 2026

Table 1. Overview of case studies. (Cont'd)

Name of initiative (URL) Lead institution(s) Brief description/mission	Type of data	Size	Database access	CDM(if applicable)	Utility	Ref(s)
GDC The University of Chicago, NCI Center for Cancer Genomics Information system for storing, analyzing, and sharing genomic and clinical data from patients with cancer	Multiple data types. Genomic, transcriptomic, clinical, and biospecimen data.	84,609 cases from 68 projects, including 69 sites	Access to GDC-controlled data through the database of Genotypes and Phenotypes.	GDC Data Model	Provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine	(16)
Million Veteran Program VA Office of Research and Development From US military veteran volunteers World's largest medical databases of blood samples and health information	Multiple data types. VA EHR, genotyping, whole-exome sequencing, and whole genome sequencing.	690,000 participants	Data currently available to VA investigators. Future data access may be granted through Request for applications, program announcements, and consortia mechanisms.	Observational Medical Outcomes Partnership CDM	For researchers at VA. Current projects include predicting the breast cancer risk for female veterans; human papillomavirus-related cancer risk; VA-Department of Energy exemplar project on cancer; correlating clinical data and genomics in early-stage or late-stage presentation of lung cancer; and progression and prognosis of multiple myeloma in US veterans.	(22)

Downloaded from http://aacrjournals.org/cancerres/article-pdf/83/8/1183/3320604/1183.pdf by guest on 06 August 2026

Table 2. Lessons learned to ensure the success of big data in oncology.

Data operability, interoperability, and quality are critical
• Adhere to published guidelines on building interoperable data sets; use standard data specifications and data collection formats, as well as common data models to produce consistent high-quality data. • Use a data-sharing taxonomy for characterizing the utility of data sets for addressing specific questions. Factors such as accessibility, data set size, data elements, associated computing environment, presence or absence of PHI, access to healthcare system source data, and whether the data are static or longitudinal determine data utility.
Reducing the time and effort required to aggregate data is essential (move to programmatic solutions when possible)
• Require work processes with cloud-based technology stacks and automated pipelines for efficient storage, aggregation, frequent updating, and analysis of data. • Integrate data aggregation into workflows. • Incorporate QA/QC throughout work processes for timely and efficient production of data sets for analysis. Constant attention to balancing activities for enhancing and maintaining data sets is critical. • Use federated systems to improve the efficiency of data aggregation. Federated systems may be required to keep up with the vast amounts of relevant healthcare data now generated.
Collect data with the intent to share from the outset
• Encourage initiatives and collaborations to foster data sharing by the research community. Early efforts have led to strategic planning in the US and European health data regulatory agencies. • Require data beyond primary clinical phenotype from EHRs, e.g., molecular, digital histopathology, DICOM, insurance claims, prescription refill, and patient-reported outcomes data. These data can potentially enhance the application of AI/ML technology in producing medical information. • Include data from important studies and provide tools for innovative analyses (such as visualization techniques for exploring various types of data and other applications) to enhance the appeal of data sharing. • Adopt well-thought-out open data-sharing models that include data privacy regulations and practices, data cycle management, and account for/control data reanalyses.

a wholly owned nonprofit subsidiary of the American Society of Clinical Oncology (ASCO). Since its founding in 2014, the CancerLinQ network platform has grown to include more than 100 participating healthcare organizations and oncology practices in the United States. As of December 2021, its database contained more than 6 million total patients, more than 2 million of which have a primary or secondary diagnosis of a malignant neoplasm. The CancerLinQ mission is to empower the oncology community to improve quality of care and patient outcomes through transformational data analytics. CancerLinQ collects comprehensive longitudinal clinical data, both structured and unstructured, from a wide variety of electronic health records (EHR) and other source systems, aggregates the data, then harmonizes, normalizes, and curates them to conform to a Common Data Model (CDM) to support queries. The data are delivered back to the contributing practices as dashboards, reports, and a suite of electronic clinical quality measures, to be used for quality improvement and clinical care. Additionally, the aggregated data undergo software-based Health Insurance Portability and Accountability Act-compliant deidentification and can then be used for data exploration and insights by practices and for discovery by the broader oncology community, including academic researchers, nonprofits/government agencies, and life sciences companies (4, 5).

Lessons learned

- Operationalizing collection, aggregation, and normalization of massive amounts of real-world oncology data (RWD) at scale requires a highly flexible cloud-based technology stack and an automated data pipeline to enable frequent updates to the aggregated data set. Additionally, the ability to integrate with all the leading information systems is critical for broad adoption, as is an open platform for developing a broad range of third-party applications.
- Integration into the oncology practice workflow is essential for physician acceptance, and there should be minimal additional data capture required of the healthcare provider. Solutions that can be used for quality reporting for value-based care arrangements or solving other practice challenges that affect revenue cycles can be enormously attractive to business owners tracking return on investment.
- Encouraging better data hygiene in physician documentation practices and greater reliance on structured data input can be a delicate negotiation with busy clinicians, who tend to have a high comfort level with dictation and free text. Wider adoption of standard data specifications like Minimal Common Oncology Data Elements (mCODE) within the EHR itself can eliminate some of this burden.
- The learning health system requires data beyond primary clinical phenotypic data from EHRs. Notable gaps include structured molecular data [predominantly somatic next-generation sequencing (NGS) reports], insurance claims data, prescription refill data, and patient-reported outcomes data. Inclusion of digital histopathology data and Digital Imaging and Communications in Medicine (DICOM) data can potentially enable the widespread application of artificial intelligence (AI)/machine-learning (ML) technologies to extract even greater information and utility.
- The recently published final rules on interoperability from the Centers for Medicare and Medicaid Services and the Office of the National Coordinator for Health Information Technology should significantly improve overall cancer data interoperability (https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi). The requirement for application programming interfaces (API) to use the modern Fast Healthcare Interoperability Resources (FHIR) standard (https://hl7.org/FHIR/) should be transformational for data exchange and the ability of patients to access their own data.

American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange

American Association for Cancer Research (AACR) Project Genomics Evidence Neoplasia Information Exchange (GENIE) is an international pan-cancer clinical-genomic registry of RWD assembled by sharing data between 19 leading academic cancer centers in an active consortium (6). The primary goal of the registry is to improve clinical decision-making, particularly in the case of rare cancers and rare variants in common cancers (7), by collecting data from nearly every patient sequenced at participating institutions.

Downloaded from http://aacrjournals.org/cancerres/article-pdf/83/8/1183/3320604/1183.pdf by guest on 06 August 2026

Through the efforts of Sage Bionetworks (https://sagebionetworks.org/) and cBioPortal for Cancer Genomics (https://genie.cbioportal.org/), the registry aggregates, harmonizes, and links clinical-grade NGS data with clinical outcomes obtained during routine medical practice from cancer patients treated at participating institutions. The consortium and its activities are driven by openness, transparency, and inclusion, ensuring that project output remains accessible to the global cancer research community for patient benefit. Details of project governance, operations, and participant sentiments have been described previously (8).

The tenth public data release occurred in July 2021 and contained the clinical-genomic sequencing results of 120,953 samples from 111,222 patients. As of December 2021, nearly 10,000 individuals had registered to use the data, and more than 500 papers had cited the registry. The first 4 public data releases are also available for analysis within the NCI GDC (https://gdc.cancer.gov/about-gdc/contributed-genomic-data-cancer-research/genie).

At the outset of the project, it was decided to harmonize existing data from each participating institution instead of agreeing to a common platform/methodology and prospectively collecting data. A significant advantage of harmonizing existing data is that you can rapidly generate large data sets. For example, the GENIE registry crossed the 100,000-patient mark in 4 short years. However, this approach has trade-offs, including some missingness across the data set; for example, if an institution does not assay for a particular gene. Additionally, there are complexities involved in harmonization and quality control (QC) for hundreds of different data sources. For example, the 10.0 public release includes data from 92 different sequencing panels and covers 1,348 unique genes.

The harmonization process begins with file preparation at each institution, where files to be transferred are mapped to prespecified formats. Generally, data transfers contain all high-quality, somatic calls, including variants of unknown significance, which have been locally reviewed by the institution. During the upload process, files are checked against a file validator by the system, and submitters are notified of issues. Sage Bionetworks processes, harmonizes, filters, and QCs the data monthly before internal release.

As highlighted throughout this text, QC is paramount and included at multiple steps throughout the data transfer process. Before initial data transfer, data providers filter any known artifacts, as well as any known germline variants. After harmonization, the data are filtered centrally for the correct date for inclusion, mutations in cis, patient retraction, and potential germline mutations. Finally, before each public release, a dedicated working group manually reviews an entire release looking for data artifacts and inconsistencies. A suite of internal tools, as well as shared algorithms, helps flag potentially problematic institutional data for review when compared with the entirety of a release (e.g., mutation frequencies, demographic distribution, etc.). The output of each review is frequently used to develop additional code and filters to help prescreen subsequent releases to the extent possible. Finally, all prior data are overwritten with each submission, ensuring that the most recent data release is as accurate and current as possible. Archival copies of all prior data releases are kept for reference and to maintain analytic integrity.

Patient protection is at the forefront of all processes. Each institution either consents patients for data sharing or provides data through an Institutional Review Board approval or waiver. Data are currently deidentified following Safe Harbor protocols, and all dates are converted to intervals from various anchor dates. Importantly, a simple click-through terms of access was implemented to protect patient identities while making data access as easy as possible (https://docs.google.com/forms/d/e/1FAIpQLScwlJ9WRmAGZ08CCg8wYo8l8bcUmsAzJ09i1MKjBNtb_dLqIw/viewform). Additionally, both explicit and implicit patient retraction processes are deployed, allowing for active or passive patient removal, respectively. Finally, internal filters have been developed to remove any potential germline mutations and/or identify single-nucleotide polymorphisms as an additional layer of patient protection.

Lessons learned

When AACR Project GENIE first launched, the initiative was as much a sociological experiment as it was a clinical research project. It was a coalition of those willing to put aside apprehensions, and it helped catalyze a cultural shift toward sharing and collective work. Quality assurance (QA)/QC is a continuously iterative process with shared responsibility between the data providers and central project administration. Each data release provides insight that is incorporated into the underlying architecture to aid with future releases.

Early in the project, QA/QC was left to the last step before a public release, which overly complicated releases and introduced delays. By adopting a monthly internal release schedule, the data quality improved, as did "on-time" deliveries, and the QA/QC process became easier. The consortium has built an extensible operational framework focusing on using existing standards whenever available, with the goal of "future-proofing" the project to the extent possible. Finally, there is a natural temporal lag built into the system. As the amount of data increases and more applications for real-time data use become apparent, the group is looking toward a shift to a fully federated system.

Project Data Sphere

Project Data Sphere (PDS) was established in 2014 to catalyze patient-focused cancer research and accelerate new therapy development. Its open-access digital library laboratory provides a secure platform for researchers to access deidentified patient-level clinical data (9, 10).

Currently, more than 2,600 authorized users on the PDS platform can access 200+ clinical trial data sets representing 200,000+ patients suffering from various tumor types, including breast, colorectal, esophageal, stomach, leukemia, lymphoma, multiple myeloma, ovarian, uterine, and prostate. To ensure researchers can realize the full potential of these data, PDS partnered with SAS Institute, which provides data mining and ML tools within the PDS environment. Research using these data sets has led to more than 135 peer-reviewed publications. Some notable examples include refs. 11–13.

To ensure success, PDS follows these principles:

- Uses an open-access model; following a simple and fast user registration, platform users can freely peruse the 200+ data sets.
- Data on the platform consist of highly annotated late-stage clinical trial data sets that are ideal for data-powered hypothesis testing.
- Users' ability to analyze data with SAS data analytic and visualization tools in the cloud or to download raw data files for analysis in their local environments maximizes accessibility and interoperability.
- The platform can adapt to evolving data opportunities. The platform started hosting solely clinical trial data. In 2020, the capabilities were expanded to host medical imaging, registry, and genomic data. This combination of characteristics has been the driver of a remarkably high ratio of publications to data sets.

Downloaded from http://aacrjournals.org/cancerres/article-pdf/83/8/1183/3320604/1183.pdf by guest on 06 August 2026

Lessons learned

Despite significant progress, numerous challenges to widespread data sharing and reuse remain. Specifically, various data providers are uncomfortable with providing, or are unable to provide, open access to their data, and prefer gatekeeper models. Reasons for this range from data privacy laws and data life cycle management (generation, release, and update) to the fear of divergent data reanalyses and competitive advantage concerns. An *ad hoc* exercise is currently required for each research project to navigate discovery, obtain permissions, access, and consolidate data. Our call to action for the research community is to conduct data generation with the expectation of open data sharing at some point in the data life cycle.

The National Cancer Institute Genomic Data-Commons

The NCI GDC (14–16) was launched in 2016 and was one of the first large-scale data-commons. A data commons collocates data with computing infrastructure and software services, tools, and applications to create a data platform for managing, harmonizing, analyzing, and sharing data sets (17). It is especially valuable for large data sets that can be challenging to manage and analyze without large-scale cloud computing infrastructure.

As of January 1, 2022, the GDC contained over 84,000 cases and 3.7 petabytes (PB) of data spanning molecular, image, and clinical data. More than 50,000 researchers use it monthly accessing more than 1 PB of data.

One of the challenges faced by the GDC was to develop an architecture that could (i) manage PB-scale genomics data; (ii) support a rich data model containing clinical, phenotypic, biospecimen, and imaging data; and (iii) provide an experience that enabled users to interactively explore the large amounts of data managed by the GDC. The GDC used a cloud-based architecture for this, initially a private cloud hosted at The University of Chicago, and later a hybrid cloud spanning The University of Chicago data center and Amazon Web Services. Through the NCI CRDC, data are made available both in Amazon Web Services and Google Cloud Platform for cloud-based applications.

The GDC manages two types of data. The first is object data, such as BAM files, which are identified by persistent globally unique identifiers (GUID). A service translates each GUID into the object's physical location, which may be in multiple locations, either in the on-premises cloud or in a public cloud. This approach allows the data to be moved or replicated without changing any of the code that references the data. The second type is structured data, such as clinical data, phenotype data, or metadata, which is stored in a database. The structured data also include metadata about each of the data objects. In this way, all data in the GDC meet the FAIR standards (18). Importantly, this architecture has enabled the GDC to scale substantially since its launch, both in number of users and amount of data. Because all the data in the GDC are available through open FAIR APIs (19), a rich set of applications has been built over the data in the GDC, both by the GDC itself and by third parties.

The use of cloud computing also provided the flexibility, scalability, and burst capability required so that the GDC could harmonize all the data submitted to it using a common set of bioinformatics pipelines (20) within a fixed time after data submission.

Lessons learned

Factors contributing to the wide use of the GDC:

- Includes data from important and interesting studies, including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET; ref. 14).
- Includes an interactive visualization to explore both molecular and clinical data, with the ability to produce and download publication-quality figures.
- All data in the GDC, comprising over 68 projects, are harmonized with respect to a CDM (the GDC data model). All data in the GDC are processed with a common uniform set of bioinformatics pipelines, making the data much simpler to understand and analyze (20).
- The GDC has an open API with a rich collection of applications built around the API (19).
- The GDC's API makes it easy to access data from the GDC in Jupyter Notebook, RStudio, and other API-based applications.

Important challenges for systems such as the GDC include:

- Reducing the time and effort to ingest new data sets. Although the GDC provides an API to upload data, the API requires understanding the GDC data model and transforming data into a format compatible with the API. This can be a time-consuming step, and the GDC has recently developed several tools to help data submitters format data correctly for uploading.
- There are currently more than 25 bioinformatic pipelines (20). These pipelines must not only be run over all submitted data in a timely fashion, but also over all the relevant data whenever any of the pipelines are updated. The GDC has developed a large-scale bioinformatics execution service for this purpose called the GDC Pipeline Automation System.
- Enhancing the functionality of the GDC while operating the GDC and improving its efficiency. As is the case for many large-scale operational systems, each year, the GDC must balance enhancing the system's functionality, adding more projects, refreshing old functionality, updating the technology stack, and improving overall efficiency.

Veterans Health Administration Clinical Data Initiative

The most extensive integrated healthcare system in the United States, the US Department of Veterans Affairs (VA) Veterans Health Administration (VHA), aggregates a large-scale data repository consisting of various modalities, including clinical, imaging, and genomic data from EHR. Although the primary reason for collecting data is for delivering healthcare to patients, these RWD have also proven essential for supporting QA, cutting-edge research, and other healthcare system needs (21).

The VHA also executes several data-related initiatives or initiatives with a data aggregation component that operate outside of the EHR. These projects include efforts by the Cooperative Studies Program (started in 1972) to capture data for clinical trials and other epidemiologic studies, the Million Veteran Program (beginning in 2011) for collecting germline sequencing data and patient-reported demographics (22), and the National Precision Oncology Program (NPOP, 2013) for collecting tumor tissue sequencing data for tailoring oncology treatment. With these critical elements in place, VHA has been able to position itself as a pioneer for using data technology to meet the operational challenges of a large healthcare enterprise, as well as to support administrative and

Downloaded from http://aacrjournals.org/cancerres/article-pdf/83/8/1183/3320604/1183.pdf by guest on 06 August 2026

research needs. However, establishing successful VA collaborations with outside entities remains challenging, mainly because of the need to maintain veterans' health data security and privacy. The following sections focus on efforts to share data with external trusted partners and third-party users that are in accordance with all VA regulatory requirements (23). Different data-sharing configurations will also be compared to show that the way data are shared greatly affects the classes of clinical questions that can be answered.

Recent data-sharing efforts

Although VHA has carried out many clinical and research data-sharing projects, most involve sharing explicitly defined static data sets specific to the project needs and with particular collaborators. In contrast, recent data-sharing efforts undertaken by VA Boston Healthcare System in the Research for Precision Oncology Program (RePOP) project aim to establish workflows and processes to enable the sharing of longitudinal VA data over more extended periods of time. RePOP, the research component of NPOP, was funded to consent NPOP patients to share their clinical, imaging, and genomic data with researchers to further advancements in cancer care. The overall workflow consists of 4 main technical components: data aggregation, deidentification, formatting, and upload. Aggregation identified relevant cohorts, determined data elements, and linked different modalities by patient identifier. Deidentification was performed using standard methods for external patient identifier generation, date obfuscation (TCGA), and DICOM header stripping for imaging [The Cancer Imaging Archive (TCIA)]. The data were then formatted and bundled according to the receiving data repository requirements and then uploaded into respective external data repositories. The initial use case for this framework was the Applied Proteogenomics OrganizationaL Learning and Outcomes (APOLLO) network (24), where clinical data were shared with the GDC (25), and imaging was shared with TCIA (26). Subsequent data transfers were made to The University of Chicago Center for Translational Data Science.

Regulatory requirements were also associated with several of the technical tasks. For aggregation, formal data requests were executed to pull data from the Corporate Data Warehouse (CDW). Internal data use agreements (DUA) were required with various imaging centers to acquire imaging data. The crosswalk between patient Medical Record Numbers and external identifiers was initially not approved to exist but eventually was allowed to remain solely on a secure VA server. This approval was crucial for longitudinal data because it enabled future data sets to link with previously shared data properly. The deidentification processes were reviewed and approved by VA Information Security Officers (ISO) and Privacy Officers (PO). For data upload, submission portals for external data repositories were reviewed and approved by ISOs and POs. DUAs between the VA and data portal administrative entities were executed according to standard VA policy.

Using this framework, the clinical, imaging, and genomic data of three related cohorts of cancer patients have been shared outside the VA: (i) consented patients to the GDC and TCIA; (ii) consented patients to The University of Chicago, and (iii) deceased patients to The University of Chicago. The first cohort is a component of the APOLLO network and consists of patients who have given consent for their data to be shared with outside entities and conforms to the data models of GDC and TCIA. The second cohort is like the first, except that the original CDW data model is preserved as much as possible. The third cohort is much larger than the first two and consists of unconsented deceased patients, where the data are mainly clinical with some imaging and genomic data. The first two cohorts have been approved to be downloaded by third-party users, but the third cohort is required to always remain in The University of Chicago environment and can be accessed only by trusted partners of the VHA.

Lessons learned

The process described in the previous section resulted in several data access configurations. The different configurations illustrate that the way the data are shared greatly affects their ability to address certain classes of clinical questions. As such, it is useful to develop a data-sharing taxonomy based on several key features to determine the advantages and disadvantages of each configuration. Data-sharing utility can be described by factors such as accessibility, data set size, data elements, associated computing environment, presence or absence of protected health information (PHI), access to healthcare system source data, and whether the data are static or longitudinal.

In turn, each of these factors affects the data utility (21). Accessibility, for example, affects what expertise can utilize the data. The size of the data set affects the level of statistical power or the amount of training data generated for ML algorithms. The types of data elements being captured fundamentally determine the queries that can be answered. The computing resources associated with a data repository impact what types of analysis can be performed (e.g., deep learning has certain minimal computing requirements). Data sets that include PHI have higher levels of fidelity to clinical care but cannot be easily shared with collaborators. Direct access to the healthcare system can support studies that require collecting patient-reported outcomes, and longitudinal, periodically refreshed data sets can support prospective studies. Finally, conclusions made using data that do not include diverse racial and ethnic populations may not be applicable to the non-included groups and care needs to be made in making such assertions.

Conclusions

Every stakeholder who touches patient data shares responsibility for delivering on the vision of harnessing the totality of available data to drive decision-making in favor of patients everywhere. The authors hope that by describing some of the emerging resources, we raise awareness and inspire generators, stewards, and consumers of healthcare data to consider the secondary use of such data at the earliest possible step. This will ensure the proper sharing of data to generate insights so that people suffering from cancer and their loved ones stand the best chance to benefit from the collective knowledge of the cancer community. One factor not addressed, but critical, is the representativeness of the data with respect to the intent-to-treat/study population. With the exception of PDS, each of the initiatives highlighted here is an example of real-world data and reflects the respective patient populations. Discrepancies between the observed and anticipated patient demographics can be attributed to numerous factors such as setting, community versus academic referral center, for example. Efforts to ensure that cancer care is as inclusive as possible, combined with more inclusive clinical trial enrollment will ultimately lead to more representative data sets in the near future.

Authors' Disclosures

S.M. Sweeney is an employee of the American Association for Cancer Research; however, this had no bearing on the review or acceptance of this manuscript. G.J.

Downloaded from http://aacrjournals.org/cancerres/article-pdf/83/8/1183/3320604/1183.pdf by guest on 06 August 2026

Davis reports other support from Abbott Laboratories outside the submitted work, being an employee of Abbott Laboratories, and also the owner of Abbott stock. S.H. Gawel is an employee of Abbott Labs. I.R. Kirsch reports personal fees from Adaptive Biotechnologies outside the submitted work. D. McGraw reports personal fees from Invitae, Datavant, the Federal All of Us Research Program, and the Centers for Disease Control and Prevention outside the submitted work. D. Milgram reports other support from the NCI and Foundation for the NIH outside the submitted work. R.S. Miller is an employee of the American Society of Clinical Oncology (ASCO). CancerLinQ is a wholly owned, nonprofit subsidiary of ASCO. T. O'Brien is a full-time employee of Pfizer. P. Robbins reports other support from Pfizer during the conduct of the study. E.H. Rubin reports other support from Merck & Co. outside the submitted work. C.C. Sigman reports other support from NCI and Foundation for the NIH outside the submitted work. No disclosures were reported by the other authors.

Acknowledgments

The authors would like to thank the Foundation for the NIH for their support and guidance, as well as Lukas Amler, Michelle Berny-Lang, Vladimir Popov, Elizabeth Pulte, Maggie Scully, Margaret Thompson, and Michael Taylor for their insight, review, and thoughtful contributions.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

Received April 15, 2022; revised July 29, 2022; accepted December 6, 2022; published first January 10, 2023.

References

1. Sweeney SM, Hamadeh HK, Abrams N, Adam SJ, Brenner S, Connors DE, et al. Challenges to using big data in cancer. Cancer Res 2023;83:1175–82.
2. Mangravite LM, Sen A, Wilbanks JT, Sage Bionetworks Governance Team. Mechanisms to govern responsible sharing of open data: a progress report. 2020. Seattle, WA: Sage Bionetworks. Available at https://sage-bionetworks.github.io/governanceGreenPaper/manuscript.pdf.
3. European Medicines Agency (EMA). Draft guideline on registry-based studies. EMA/502388/2020. 2020.
4. 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;32:2373–9.
5. Potter D, Brothers R, Kolacevski A, Koskimaki JE, McNutt A, Miller RS, et al. Development of CancerLinQ, a health information learning platform from multiple electronic health record systems to support improved quality of care. JCO Clin Cancer Inform 2020;4:929–37.
6. AACR Project GENIE Consortium. AACR project GENIE: powering precision medicine through an international consortium. Cancer Discov 2017;7:818–31.
7. Smyth LM, Zhou Q, Nguyen B, Yu C, Lepisto EM, Arnedos M, et al. Characteristics and outcome of AKT1 (E17K)-mutant breast cancer defined through AACR project GENIE, a clinicogenomic registry. Cancer Discov 2020;10:526–35.
8. Micheel CM, Sweeney SM, LeNoue-Newton ML, André F, Bedard PL, Guinney J, et al. American association for cancer research project genomics evidence neoplasia information exchange: from inception to first data release and beyond-lessons learned and member institutions' perspectives. JCO Clin Cancer Inform 2018;2:1–14.
9. Green AK, Reeder-Hayes KE, Corty RW, Basch E, Milowsky MI, Dusetzina SB, et al. The project data sphere initiative: accelerating cancer research by sharing data. Oncologist 2015;20:464–e20.
10. Bertagnolli MM, Sartor O, Chabner BA, Rothenberg ML, Khozin S, Hugh-Jones C, et al. Advantages of a truly open-access data-sharing model. N Engl J Med 2017;376:1178–81.
11. Wilkerson J, Abdallah K, Hugh-Jones C, Curt G, Rothenberg M, Simantov R, et al. Estimation of tumour regression and growth rates during treatment in patients with advanced prostate cancer: a retrospective analysis. Lancet Oncol 2017;18:143–54.
12. Guinney J, Wang T, Laajala TD, Winner KK, Bare JC, Neto EC, et al. Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. Lancet Oncol 2017;18:132–42.
13. Seyednasrollah F, Koestler DC, Wang T, Piccolo SR, Vega R, Greiner R, et al. A DREAM challenge to build prediction models for short-term discontinuation of docetaxel in metastatic castration-resistant prostate cancer. JCO Clin Cancer Inform 2017;1:1–15.
14. 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;375:1109–12.
15. Heath AP, Ferretti V, Agrawal S, An M, Angelakos JC, Arya R, et al. The NCI genomic data commons. Nat Genet 2021;53:257–62.
16. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI genomic data commons as an engine for precision medicine. Blood 2017;130:453–9.
17. Grossman RL. Data lakes, clouds, and commons: a review of platforms for analyzing and sharing genomic data. Trends Genet 2019;35:223–34.
18. 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.
19. 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;77:e15–e8.
20. Zhang Z, Hernandez K, Savage J, Li S, Miller D, Agrawal S, et al. Uniform genomic data analysis in the NCI genomic data commons. Nat Commun 2021;12:1226.
21. Fihn SD, Francis J, Clancy C, Nielson C, Nelson K, Rumsfeld J, et al. Insights from advanced analytics at the veterans health administration. Health Aff (Millwood) 2014;33:1203–11.
22. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million veteran program: a mega-biobank to study genetic influences on health and disease. J Clin Epidemiol 2016;70:214–23.
23. Do N, Grossman R, Feldman T, Fillmore N, Elbers D, Tuck D, et al. The veterans precision oncology data commons: transforming VA data into a national resource for research in precision oncology. Semin Oncol 2019;46:314–20.
24. Fiore LD, Rodriguez H, Shriver CD. Collaboration to accelerate proteogenomics cancer care: the department of veterans affairs, department of defense, and the national cancer institute's applied proteogenomics organizational learning and outcomes (APOLLO) network. Clin Pharmacol Ther 2017;101:619–21.
25. Contreras JL, Knoppers BM. The genomic commons. Annu Rev Genomics Hum Genet 2018;19:429–53.
26. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045–57.

Downloaded from http://aacrjournals.org/cancerres/article-pdf/83/8/1183/3320604/1183.pdf by guest on 06 August 2026