THE UNIVERSITY OF CHICAGO

CENTER FOR TRANSLATIONAL DATA SCIENCE

01

2022 ANNUAL REPORT

2022

CENTER FOR TRANSLATIONAL DATA SCIENCE

THE UNIVERSITY OF
CHICAGO

CTDS ANNUAL REPORT

TABLE OF CONTENTS

CENTER FOR TRANSLATIONAL DATA SCIENCE

MESSAGE FROM **THE DIRECTOR**

Robert L. Grossman Ph.D.

Dear Colleagues,

The Center for Translational Data Science (CTDS) had a productive year in 2022 and made important progress on our core mission of harnessing data science to advance the fields of biology, medicine, health care, and the environment. You can see our 2022 research publications on our CTDS web site. In terms of our goal of accelerating research around the world through secure data sharing and cloud-based data exploration and analysis, we added new capabilities to our existing projects, brought new projects online, and added new features to the core underlying Gen3 software platform.

Here are a just a few highlights of 2022:

During 2022, our Veterans Affairs Data Commons supporting the VA Million Veterans Program completed its beta launch, as did our Biomedical Research Hub data mesh.

During an average month in 2022, over 60,000 researchers used the Genomic Data Commons (GDC) and accessed or downloaded over 2 Petabytes of data. To date, PubMed lists over 500 research papers that are based upon data from the GDC.

We also launched the Gen3 Community Forum, which has a virtual meeting every other month to support the Gen3 community around the world.

Our research focus moved from data commons to data meshes (aka data ecosystems) that consist of multiple data commons and data repositories, cloud-based computational resources, and other cloud-based resources that interoperate using a small set of software services called framework services. We launched new versions of two of our data meshes: the HEAL Data Platform and the Biomedical Research Hub.

The progress in this report is due to the hard work of dozens of CTDS researchers and staff as well as our many other collaborators and contributors from around the world. It is also due to our sponsors and the enormous opportunity they see in how data commons, and Gen3 in particular, can have a positive impact on research. Thanks also to the readers of this report for your interest in CTDS and our mission of applying data science to hard questions in biomedical research. We look forward to sharing our updates from 2023 with you next year!

ROBERT GROSSMAN, PHD
Center for Translational Data Science Director

CENTER FOR TRANSLATIONAL DATA SCIENCE **AT A GLANCE**

From Year 2022

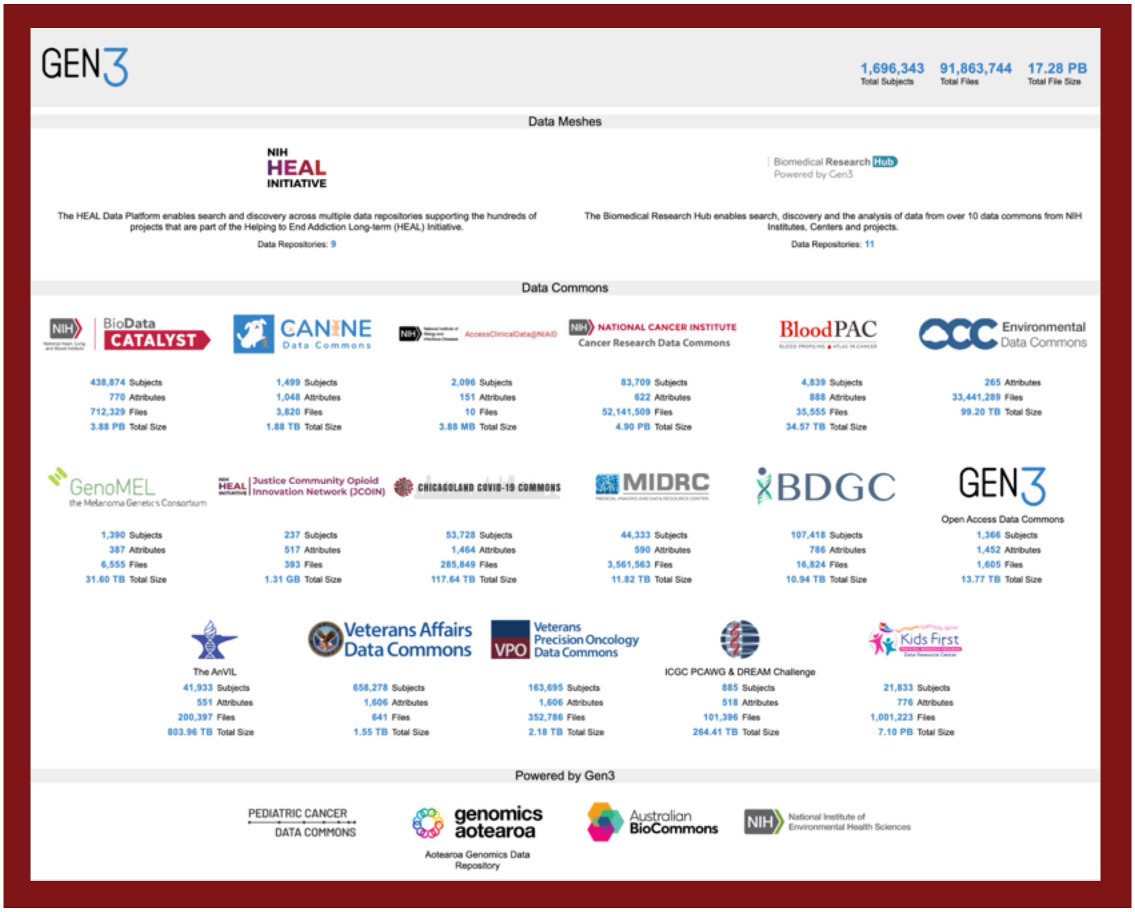
ACTIVITY OF CTDS

The Center for Translational Data Science at the University of Chicago is developing the discipline of data science and its applications to problems in biology, medicine, healthcare and the environment.

We create and operate large scale data platforms to support research in topics of societal interest, including cancer, cardiovascular disease, inflammatory bowel disease (IBD), birth defects, veterans' health, pain management, opioid use disorder, and environmental science. We also develop new machine learning and AI algorithms over the data in our platforms.

Today with our partners, we operate a data ecosystem comprising over 20 data commons that make over 17 PB of data available to the research community from nearly 1.7M patients.

We provide access to this data via secure and compliant workspaces, while protecting patient privacy. These are all based on the open-source Gen3 data platform, that includes data commons, framework and mesh services, and workspaces.


GEN3
NIH HEAL INITIATIVE
Biomedical Research Hub
Powered by Gen3
NIH BioData CATALYST
CANINE Data Commons
NIH AccessClinicalData@NIAID
NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons
BloodPAC
Environmental Data Commons
GenoMEL the Melanoma Genetics Consortium
HEAL Justice Community Opioid Innovation Network (JCOIN)
CHICAGOLAND COVID-19 COMMONS
MIDRC
BDGC
GEN3
Veterans Affairs Data Commons
VPO Veterans Precision Oncology Data Commons
Kids First
PEDIATRIC CANCER DATA COMMONS
genomics aotearoa
Australian BioCommons
NIH National Institute of Environmental Health Sciences

1,696,343 Total Subjects
91,863,744 Total Files
17.28 PB Total File Size

Data Meshes

The HEAL Data Platform enables search and discovery across multiple data repositories supporting the hundreds of projects that are part of the Helping to End Addiction Long-term (HEAL) Initiative.

Data Repositories: 9

The Biomedical Research Hub enables search, discovery and the analysis of data from over 10 data commons from NIH Institutes, Centers and projects.

Data Repositories: 11

Data Commons

438,874 Subjects
770 Attributes
712,329 Files
3.88 PB Total Size

1,499 Subjects
1,048 Attributes
3,820 Files
1.88 TB Total Size

2,096 Subjects
151 Attributes
10 Files
3.88 MB Total Size

83,709 Subjects
622 Attributes
52,141,509 Files
4.90 PB Total Size

4,839 Subjects
888 Attributes
35,555 Files
34.57 TB Total Size

265 Attributes
33,441,289 Files
99.20 TB Total Size

1,390 Subjects
387 Attributes
6,555 Files
31.60 TB Total Size

237 Subjects
517 Attributes
393 Files
1.31 GB Total Size

53,728 Subjects
1,464 Attributes
285,849 Files
117.64 TB Total Size

44,333 Subjects
590 Attributes
3,561,563 Files
11.82 TB Total Size

107,418 Subjects
786 Attributes
16,824 Files
10.94 TB Total Size

Open Access Data Commons
1,366 Subjects
1,452 Attributes
1,605 Files
13.77 TB Total Size

The AnVIL
41,933 Subjects
551 Attributes
200,397 Files
803.96 TB Total Size

658,278 Subjects
1,606 Attributes
641 Files
1.55 TB Total Size

163,695 Subjects
1,606 Attributes
352,786 Files
2.18 TB Total Size

ICGC PCAWG & DREAM Challenge
885 Subjects
518 Attributes
101,396 Files
264.41 TB Total Size

21,833 Subjects
776 Attributes
1,001,223 Files
7.10 PB Total Size

Powered by Gen3

Aotearoa Genomics Data Repository

Gen3 Data Commons are used by a long list of sponsors in order to enable secure and compliant data sharing and analysis.

HIGHLIGHTS **FROM SELECTED PROJECTS**

From Year 2022

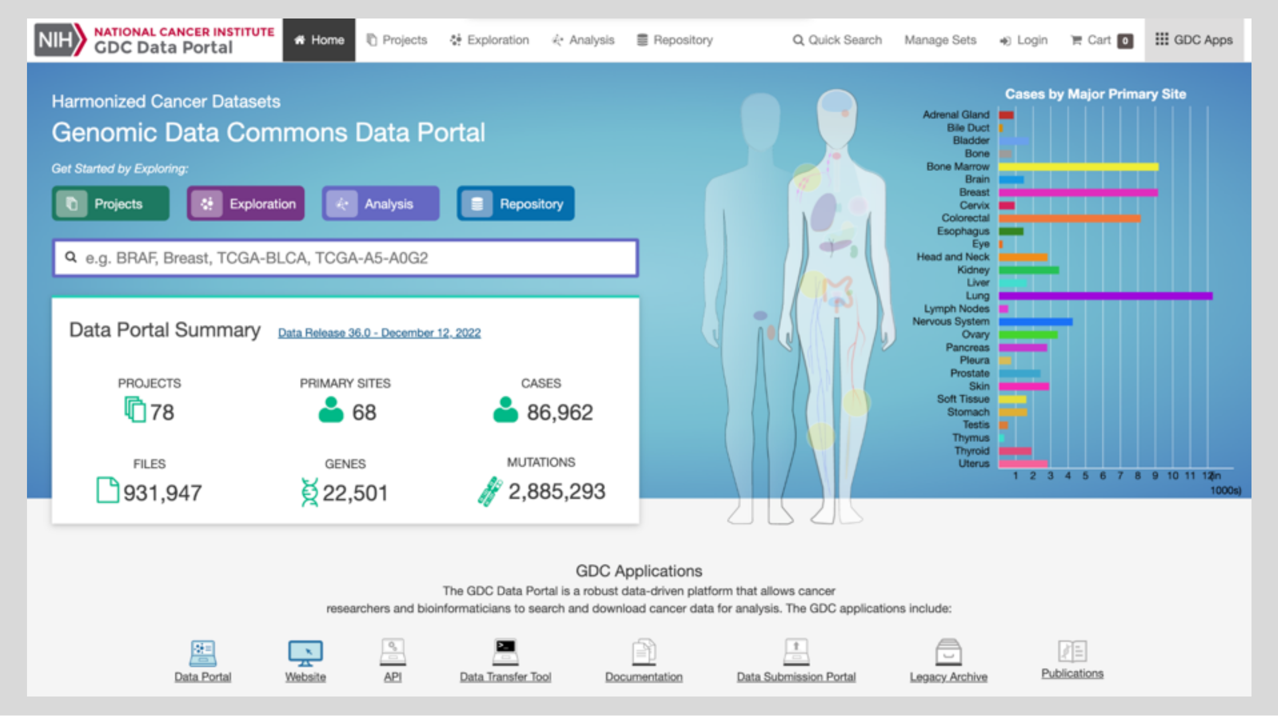
GEONOMIC DATA COMMONS

NCI's Genomic Data Commons (GDC) is the largest commons managed by CTDS and provides the cancer research community with a harmonized data repository and cancer knowledge base that enables data sharing across cancer genomic studies in support of precision medicine.

The GDC supports several cancer genome programs including The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), Clinical Proteomic Tumor Analysis Consortium (CPTAC), Multiple Myeloma Research Foundation (MMRF), and many others.

It includes clinical metadata and a range of raw and derived files from a variety of experimental strategies including whole genome sequencing (WGS), RNA-Seq, whole exome sequencing (WXS), miRNA-Seq, slide images, and many others.

In 2022, the GDC team made a major data release updating over 806k files and 875TB of data to use an updated gene model (gencode v36) across all data types. The new gene model significantly improves that data quality and includes multiple new gene types, 22% more exons, and 17% more transcripts. In 2022 the GDC averaged over 78,000 unique visitors per month.



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Data Portal Summary Data Release 36.0 - December 12, 2022

PROJECTS	PRIMARY SITES	CASES
78	68	86,962
FILES	**GENES**	**MUTATIONS**
931,947	22,501	2,885,293

GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

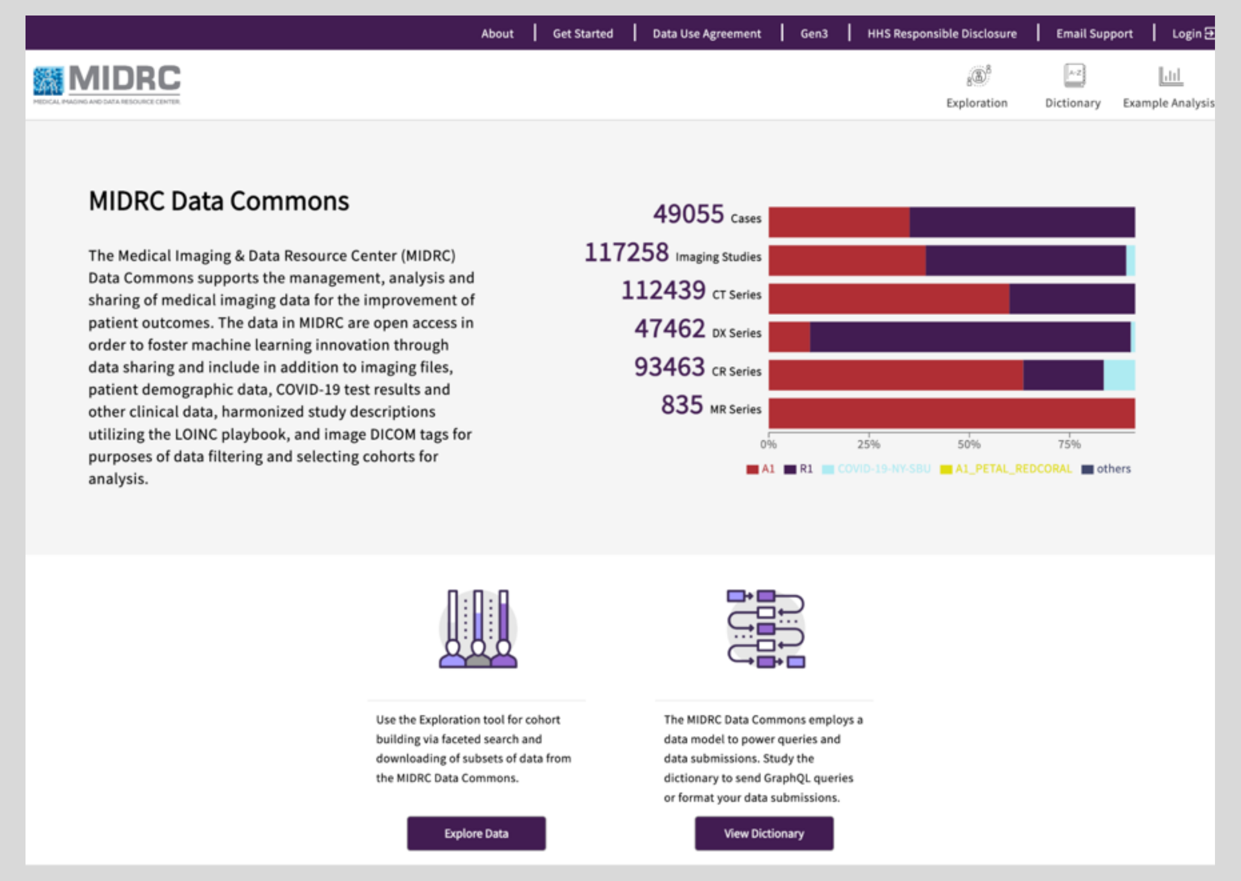
The NCI Genomic Data Commons provides harmonized cancer genomic and clinical data to the research community.

MEDICAL IMAGING AND DATA RESOURCE CENTER

The Medical Imaging and Data Resource Center (MIDRC) is a multi-institutional collaborative initiative driven by the medical imaging community and is aimed at accelerating the transfer of knowledge and innovation in the current COVID-19 pandemic and beyond. MIDRC is funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) and hosted at the University of Chicago.

It is co-led by the American College of Radiology® (ACR®), the Radiological Society of North America (RSNA), and the American Association of Physicists in Medicine (AAPM). The aim of MIDRC is to foster machine learning innovation via data sharing through rapid and flexible collection, curation and harmonization, analysis, and dissemination of imaging and associated clinical data by providing researchers with unparalleled resources in the fight against COVID-19. In just two short years, MIDRC has ingested over 300,000 imaging studies and released over 125,000 imaging studies to the public to advance the development of ML/AI for COVID19. In 2022, the MIDRC team also launched an online DICOM viewer integrated with our data portal (data.midrc.org).

This allows a user to conveniently review the DICOM images associated with each study when building cohorts. DICOM is the standard for the communication and management of medical imaging information and related data. MIDRC also provides various resources to AI investigators including a metrology decision tree and a bias-awareness tool.



MIDRC Data Commons

The Medical Imaging & Data Resource Center (MIDRC) Data Commons supports the management, analysis and sharing of medical imaging data for the improvement of patient outcomes. The data in MIDRC are open access in order to foster machine learning innovation through data sharing and include in addition to imaging files, patient demographic data, COVID-19 test results and other clinical data, harmonized study descriptions utilizing the LOINC playbook, and image DICOM tags for purposes of data filtering and selecting cohorts for analysis.

Use the Exploration tool for cohort building via faceted search and downloading of subsets of data from the MIDRC Data Commons.

The MIDRC Data Commons employs a data model to power queries and data submissions. Study the dictionary to send GraphQL queries or format your data submissions.

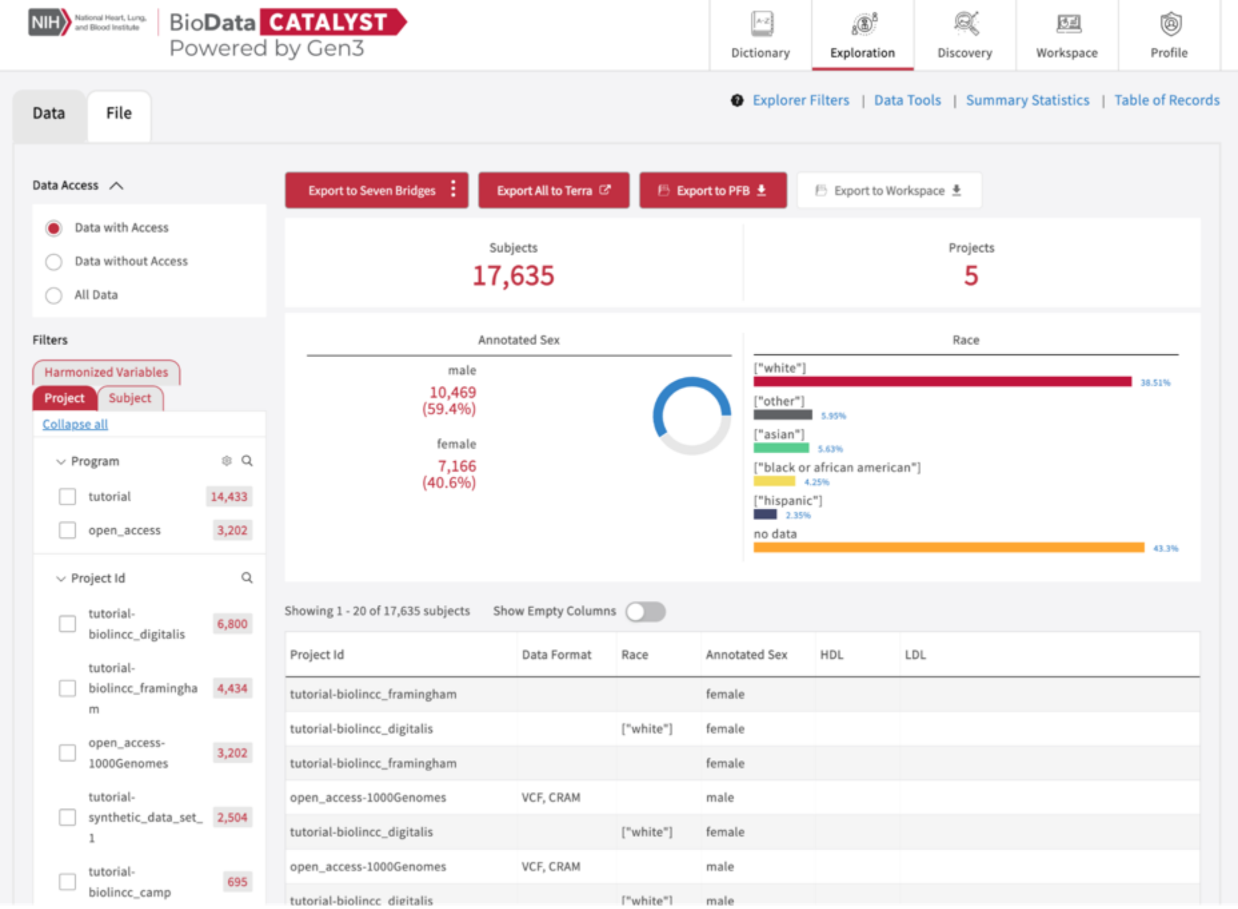
The MIDRC Data Commons provides open access to de-identified, curated, and diverse medical images of COVID-19 patients to all AI researchers. It plans to expand to other medical imaging modalities and diseases in the future.

BIODATA CATALYST

The NHLBI BioData Catalyst program is a cloud-based ecosystem providing tools, applications, and workflows in secure workspaces. By increasing access to NHLBI datasets and innovative data analysis capabilities, BioData Catalyst accelerates efficient biomedical research that drives discovery and scientific advancement, leading to novel diagnostic tools, therapeutics, and prevention strategies for heart, lung, blood, and sleep disorders.

The BioData Catalyst Gen3 Platform provides data commons services through authentication/authorization, object file indexing, interactive data search and export, and analytical workspaces services. Partner organizations and approved researchers can search and access hosted genomic and phenotypic data, and export selected cohorts to analytical workspaces in a scalable, reproducible, and secure manner.

Accomplishments in 2022 included new datasets (e.g. data from TOPMed, COVID-19 related data, and the Pediatric Cardiac Genomics Consortium) and new functionality around the Portable Format for Bioinformatics (PFB), which is a file format used to share data with other systems from Gen3.



● Data with Access
○ Data without Access
○ All Data

Program

□	tutorial	14,433
□	open_access	3,202

Project Id

□	tutorial-biolincc_digitalis	6,800
□	tutorial-biolincc_framingham	4,434
□	open_access-1000Genomes	3,202
□	tutorial-synthetic_data_set_1	2,504
□	tutorial-biolincc_camp	695

Subjects 17,635

Projects 5

male 10,469 (59.4%)

female 7,166 (40.6%)

Showing 1 - 20 of 17,635 subjects   Show Empty Columns

Project Id	Data Format	Race	Annotated Sex	HDL	LDL
tutorial-biolincc_framingham			female		
tutorial-biolincc_digitalis		["white"]	female		
tutorial-biolincc_framingham			female		
open_access-1000Genomes	VCF, CRAM		male		
tutorial-biolincc_digitalis		["white"]	female		
open_access-1000Genomes	VCF, CRAM		male		
tutorial-biolincc_digitalis		["white"]	male		

The BioData Catalyst provides access to data from programs across the National Heart Lung and Blood Institute (NHLBI). Data can be explored and analyzed within the system.

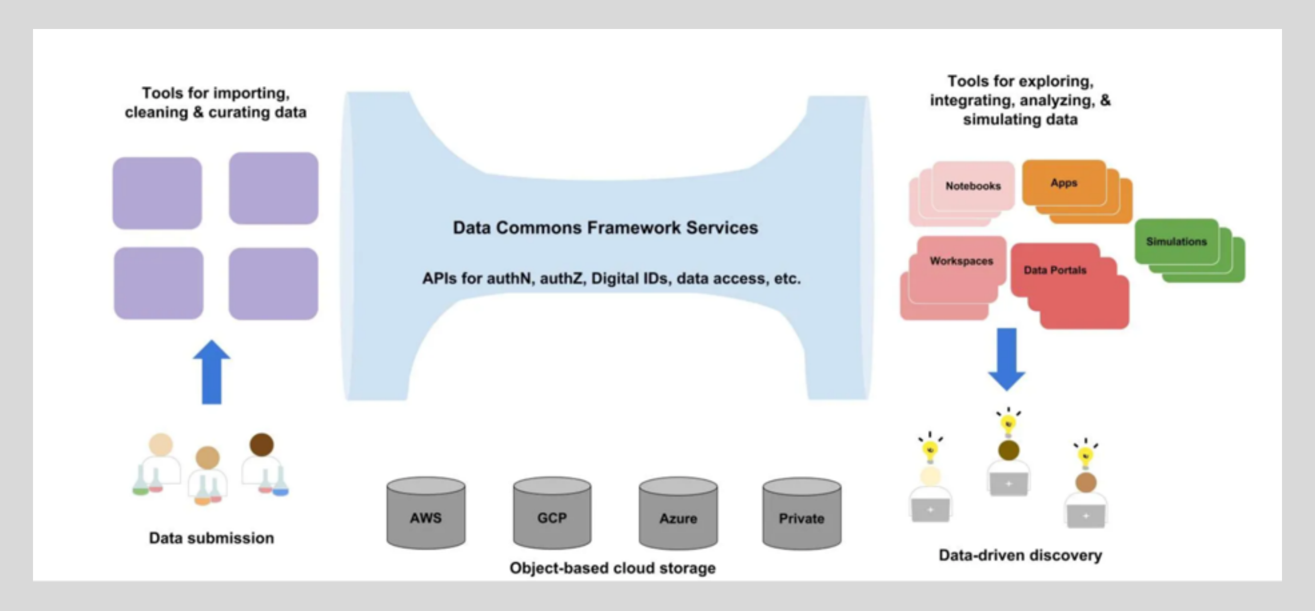
DATA COMMONS FRAMEWORK

The National Cancer Institute's Data Commons Framework Services (NCI DCFS) is a set of software services designed to make it easier to develop, operate, and interoperate data commons, data clouds, knowledge bases, and other resources for managing, analyzing, and sharing cancer research data that are part of the Cancer Research Data Commons (CRDC). The NCI DCFS supports making data Findable, Accessible, Interoperable, and Reusable (FAIR).

Data objects are assigned GUID (Global Unique Identification number) and can be stored in one or more private and public clouds and accessed using DCF services.

Structured data can be incorporated using data models and enriched with controlled vocabularies and ontologies. Gen3 includes authentication and authorization services so that controlled-access data can be handled securely. Gen3 DCF services also include the ability to define, import, and query against a data model.

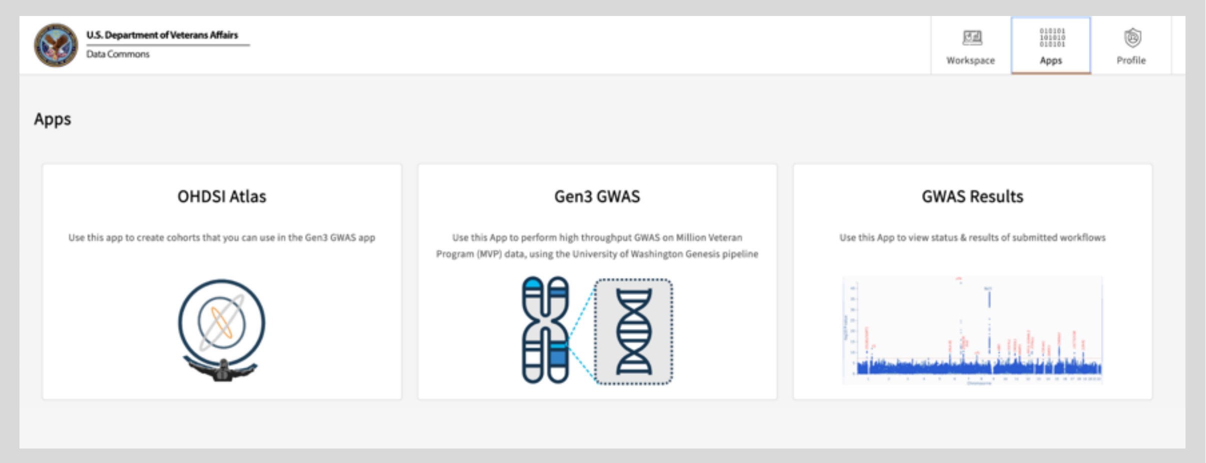
In 2022, DCF released over 50M files of data across all the nodes in the Cancer Research Data Commons. This is an increase of 2800% in number of files over 2021. These data will now be able to be analyzed in the Cloud Resources.


Tools for importing, cleaning & curating data
Data submission
Data Commons Framework Services
APIs for authN, authZ, Digital IDs, data access, etc.
AWS
GCP
Azure
Private
Object-based cloud storage
Tools for exploring, integrating, analyzing, & simulating data
Notebooks
Apps
Simulations
Workspaces
Data Portals
Data-driven discovery

The Data Commons Framework plays a central role in connecting the different resources within the Cancer Research Data Commons.

U.S. DEPARTMENT OF VETERANS AFFAIRS DATA COMMONS

The VA Data Commons supports the research and analysis of medical and genomic data from US military Veterans. It aims to accelerate scientific discovery and development of therapies, diagnostic tests, and other technologies for improving the lives of Veterans and beyond. The data commons features GWAS analyses using genetic data from veterans with clinical variables harmonized to the OMOP Common Data Model.



Apps

OHDSI Atlas

Use this app to create cohorts that you can use in the Gen3 GWAS app

Gen3 GWAS

Use this App to perform high throughput GWAS on Million Veteran Program (MVP) data, using the University of Washington Genesis pipeline

GWAS Results

Use this App to view status & results of submitted workflows

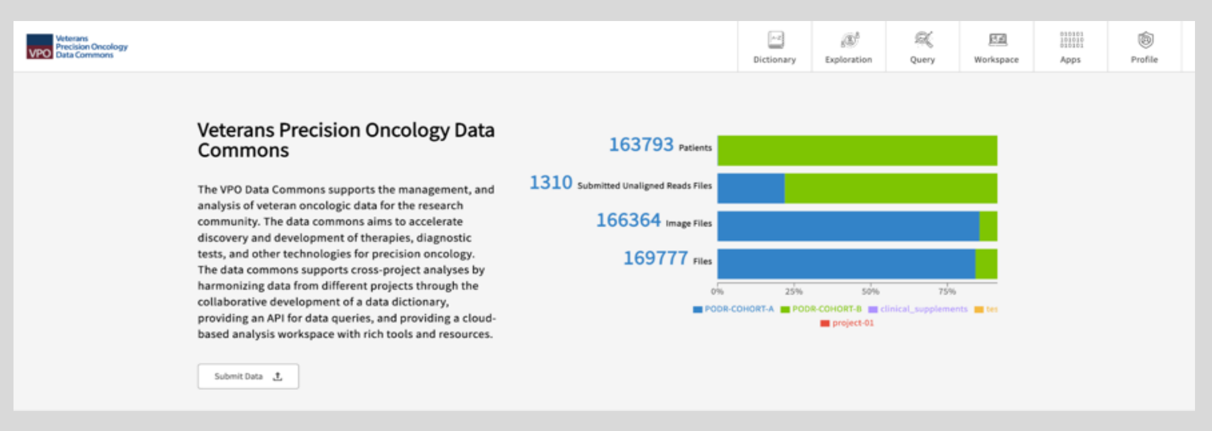
The Veterans Administration Data Commons unites data from across the VA to enable research and to improve the lives of Veterans and all Americans through health care discovery and innovation.

VETERANS PRECISIONS ONCOLOGY DATA COMMONS

The VPODC supports the management and analysis of Veteran oncology data for the research community. This data commons aims to accelerate discovery and development of therapies, diagnostic tests, and other technologies for precision oncology.

It supports cross-project analyses by data harmonization through the collaborative development of a data dictionary, providing an API for data queries, and providing an analysis workspace with rich tools and resources.

New accomplishments in 2022 included performing research and writing a manuscript on molecular signatures for prostate cancer.



Veterans Precision Oncology Data Commons

The VPO Data Commons supports the management, and analysis of veteran oncologic data for the research community. The data commons aims to accelerate discovery and development of therapies, diagnostic tests, and other technologies for precision oncology. The data commons supports cross-project analyses by harmonizing data from different projects through the collaborative development of a data dictionary, providing an API for data queries, and providing a cloud-based analysis workspace with rich tools and resources.

The Veterans Precisions Oncology Data Commons distributes clinical and genomic oncology data to researchers with the goal of improving veterans health.

HEAL DATA PLATFORM

The HEAL Platform is designed to provide a secure environment for discovery and analysis of results and data resulting from NIH HEAL-funded studies. The Platform represents a data ecosystem, or mesh, that aggregates and presents data from multiple resources to make data discovery and access easy for users. The mesh provides a way to search and query over study metadata and diverse data types, generated by different projects and organizations and stored across multiple secure repositories.

The HEAL Platform also offers a secure and cost-effective cloud-computing environment for data analysis, enabling collaborative research and development of new analytical tools. New workflows and results of analyses can be shared with the HEAL community to enable collaborative, high-impact publications with insights that address the opioid crisis.

Milestones in 2022 included launching the platform to the HEAL Initiative researcher community, establishing the ability for investigators to register their study on the Platform, providing mechanisms for investigators to submit information about their studies (metadata), integration of NIH STRIDES as a payment model for compute, and other mesh service improvements to allow for data sharing across data repositories and commons.

The Helping to End Addiction Long-term Initiative, or NIH HEAL Initiative®, is an aggressive, trans-agency effort to speed scientific solutions to stem the national opioid public health crisis. The HEAL Data Platform is funded by the NIH HEAL Initiative. NIH HEAL Initiative and Helping to End Addiction Long-term are service marks of the U.S. Department of Health and Human Services.

NEW PROJECTS

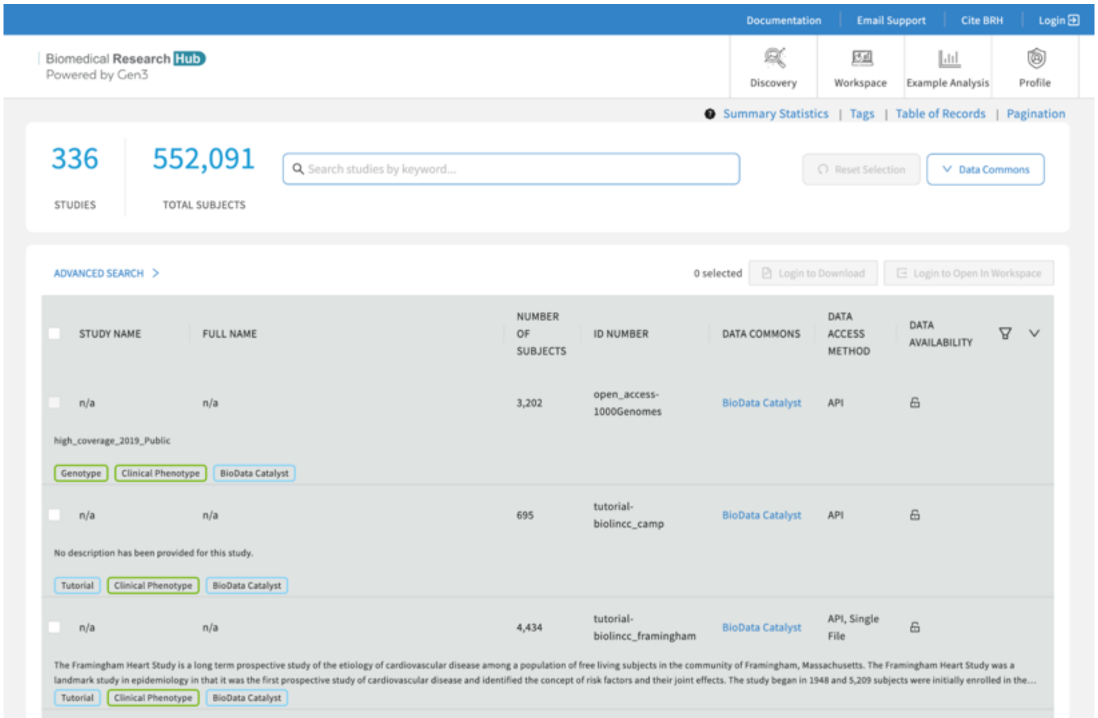
From Year 2022

BIOMEDICAL RESEARCH HUB

The Biomedical Research Hub (BRH) is a cloud-based federated system for managing, analyzing, and sharing patient data for research purposes, while allowing each resource sharing patient data to operate their component based upon their own governance rules. BRH currently interoperates with 11 separate Data Commons.

BRH framework services include authentication and authorization; services for generating and assessing findable, accessible, interoperable, and reusable (FAIR) data; and services for importing and exporting bulk clinical data. BRH includes workspaces that can access and analyze data from one or more of the data resources in the BRH.

Milestones in 2022 included beta testing, integration of NIH STRIDES as a payment model for compute, and other mesh service improvements to allow for data sharing across data repositories or commons.



336 STUDIES

552,091 TOTAL SUBJECTS

0 selected

STUDY NAME	FULL NAME	NUMBER OF SUBJECTS	ID NUMBER	DATA COMMONS	DATA ACCESS METHOD	DATA AVAILABILITY
n/a	n/a	3,202	open_access-1000Genomes	BioData Catalyst	API	🔓
n/a	n/a	695	tutorial-biolincc_camp	BioData Catalyst	API	🔓
n/a	n/a	4,434	tutorial-biolincc_framingham	BioData Catalyst	API, Single File	🔓

high_coverage_2019_Public

Genotype | Clinical Phenotype | BioData Catalyst

No description has been provided for this study.

Tutorial | Clinical Phenotype | BioData Catalyst

The Framingham Heart Study is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. The Framingham Heart Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified the concept of risk factors and their joint effects. The study began in 1948 and 5,209 subjects were initially enrolled in the...

Tutorial | Clinical Phenotype | BioData Catalyst

The Biomedical Research Hub (BRH) brings data together in one place from across several NIH systems in order to accelerate analysis and discovery.

ELWAZI OPEN SCIENCE DATA PLATFORM FOR AFRICA

eLwazi is an African-led open data science platform providing an interactive environment to apply data science techniques to diverse datasets for novel health discoveries. eLwazi provides a flexible, scalable, open data science platform for the DSI-Africa consortium to find and access data, select tools and workflows, and run analyses on a choice of computing environments, all through easy to use workspaces. Gen3 is an approved system for projects in the consortium to share data.

The African-led eLwazi consortium aims to bring together biomedical data from across Africa.

INTERNATIONAL ALLIANCE FOR CANCER EARLY DETECTION

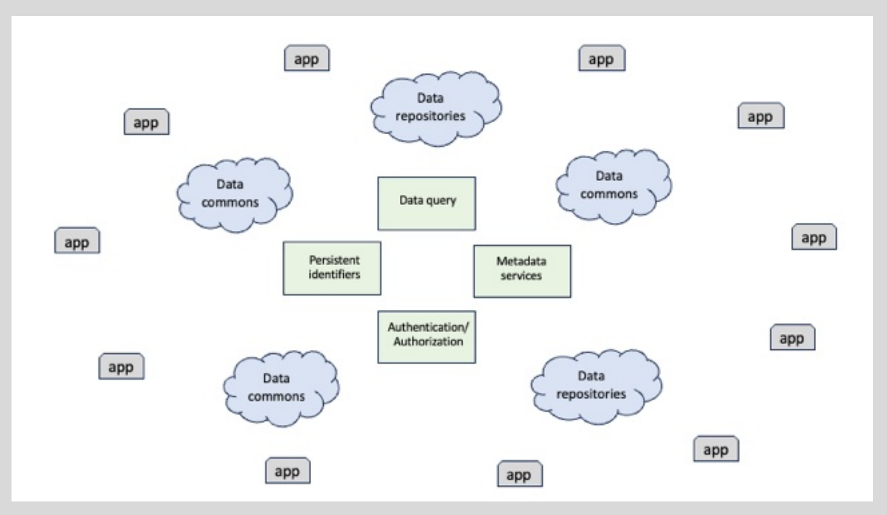
The International Alliance for Cancer Early Detection (ACED) is uniting world leading researchers to tackle the biggest challenges in early detection, an important area of unmet clinical need. Scientists in the Alliance are working together at the forefront of technological innovation to translate research into realistic ways to improve cancer diagnosis, which can be implemented into health systems and meaningfully benefit people with cancer. Gen3 will be used to support data sharing and analysis within the alliance.

The International Alliance for Cancer Early Detection (ACED) will bring together data from around the world to study and diagnose cancer at its earliest stages.

NEW **TECHNOLOGIES**

DATA MESHES

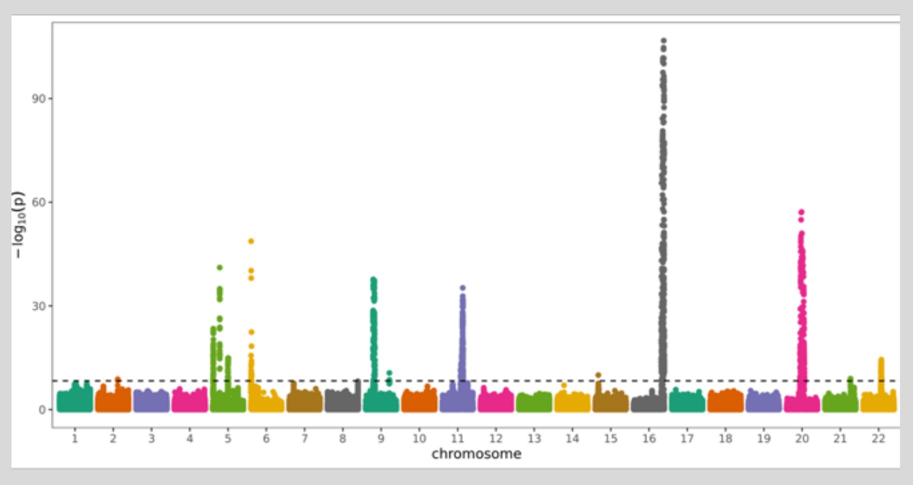
In a data mesh or data ecosystem, data commons are able to interoperate with each other and applications are able to access data and services from multiple data commons. CTDS released multiple improvements to mesh data services in 2022 including important updates to two meshes: the Biomedical Research Hub and the HEAL Platform. Gen3 mesh services (also known as data framework services) are also used to support the Cancer Research Data Commons.



Data commons and repositories are linked together by core data mesh services such as data query, persistent identifiers, metadata services, and authentication/authorization. An ecosystem of applications is able to interact with all linked data resources.

GENOME-WIDE ASSOCIATION STUDIES (GWAS)

Genome-Wide Association Studies (GWAS) Analysis is a key tool for identifying genes or genomic regions that are related to health and disease. In 2022, CTDS released a fully automated GWAS pipeline for analyzing harmonized data within the Veterans Administration Data Commons. Our pipeline is run using Argo, which is a Kubernetes-native workflow engine. Automated and reproducible bioinformatics pipelines such as this are critical to meet FAIR data standards (Finadable, Accessible, Interoperable, and Reusable).



This output of a GWAS analysis (a Manhattan Plot) shows which regions of the genome are correlated with specific diseases or phenotypes of interest.

NEW **PUBLICATIONS**

Arnovitz, S., Mathur, P., Tracy, M., Mohsin, A., Mondal, S., Quandt, J., Hernandez, K. M., Khazaie, K., Dose, M., Emmanuel, A. O., & Gounari, F. (2022). Tcf-1 promotes genomic instability and T cell transformation in response to aberrant β-catenin activation. Proceedings of the National Academy of Sciences of the United States of America, 119(32), e2201493119. https://doi.org/10.1073/pnas.2201493119

Barnes, C., Bajracharya, B., Cannalte, M., Gowani, Z., Haley, W., Kass-Hout, T., Hernandez, K., Ingram, M., Juvvala, H. P., Kuffel, G., Martinov, P., Maxwell, J. M., McCann, J., Malhotra, A., Metoki-Shlubsky, N., Meyer, C., Paredes, A., Qureshi, J., Ritter, X., Schumm, P., … Grossman, R. L. (2022). The Biomedical Research Hub: a federated platform for patient research data. Journal of the American Medical Informatics Association : JAMIA, 29(4), 619–625. https://doi.org/10.1093/jamia/ocab247

Jochum, M., Lee, M.D., Curry, K. et al. (2022) Analysis of bronchoalveolar lavage fluid metatranscriptomes among patients with COVID-19 disease. Sci Rep 12, 21125. https://doi.org/10.1038/s41598-022-25463-0

Kuang, X., Wang, F., Hernandez, K. M., Zhang, Z., & Grossman, R. L. (2022). Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN. Scientific reports, 12(1), 2427. https://doi.org/10.1038/s41598-022-06449-4

O'Hara, T., Saldanha, A., Trunnell, M., Grossman, R. L., Hota, B., & Frankenberger, C. (2022). Economical Utilization of Health Information with Learning Healthcare System Data Commons. Perspectives in health information management, 19(Spring), 1d.

Saravia-Butler A. M., Schisler J. C. , Taylor D., Beheshti A., Butler D., Meydan C., Foox J., Hernandez K., Mozsary C., Mason C. E. , Meller R. (2022) Host transcriptional responses in nasal swabs identify potential SARS-CoV-2 infection in PCR negative patients. iScience, 25(11):105310. https://doi.org/10.1016/j.isci.2022.105310

Schatz, M. C., Philippakis, A. A., Afgan, E., Banks, E., Carey, V. J., Carroll, R. J., Culotti, A., Ellrott, K., Goecks, J., Grossman, R. L., Hall, I. M., Hansen, K. D., Lawson, J., Leek, J. T., Luria, A. O., Mosher, S., Morgan, M., Nekrutenko, A., O'Connor, B. D., Osborn, K., … Wuichet, K. (2022). Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. Cell genomics, 2(1), 100085. https://doi.org/10.1016/j.xgen.2021.100085

Schumm, L. P., Giurcanu, M. C., Locey, K. J., Ortega, J. C., Zhang, Z., & Grossman, R. L. (2022). Racial and ethnic disparities in the observed COVID-19 case fatality rate among the U.S. population. Annals of epidemiology, 74, 118–124. https://doi.org/10.1016/j.annepidem.2022.07.010

Sheffield, N. C., Bonazzi, V. R., Bourne, P. E., Burdett, T., Clark, T., Grossman, R. L., Spjuth, O., & Yates, A. D (2022). From biomedical cloud platforms to microservices: next steps in FAIR data and analysis. Scientific data, 9(1), 553. https://doi.org/10.1038/s41597-022-01619-5

Shi, C., Babiker, N., Urbanek, J. K., Grossman, R. L., Huisingh-Scheetz, M., & Rzhetsky, A. (2022). Free-living wrist and hip accelerometry forecast cognitive decline among older adults without dementia over 1- or 5-years in two distinct observational cohorts. npj aging, 8(1), 7. https://doi.org/10.1038/s41514-022-00087-w

Wcisel, D. J., Dornburg, A., McConnell, S. C., Hernandez, K. M., Andrade, J., de Jong, J. L. O., Litman, G. W., & Yoder, J. A. (2022). A highly diverse set of novel immunoglobulin-like transcript (NILT) genes in zebrafish indicates a wide range of functions with complex relationships to mammalian receptors. Immunogenetics, 10.1007/s00251-022-01270-9. Advance online publication. https://doi.org/10.1007/s00251-022-01270-9.

RECENT **HIRES**

AARTI VENKAT

Aarti Venkat, Ph.D. is the Director of Clinical Informatics in the CTDS and Assistant Professor of Medicine at the University of Chicago.

Her research is broadly focused on oncology, specifically i) developing bioinformatics and machine learning methods for multimodal cancer data and ii) building data meshes, systems that researchers can use to perform federated search and learning. She is Co-PI for the Biomedical Research Hub, a leading example of a data mesh for biomedical research.

She earned her Ph.D. in Human Genetics at the University of Chicago. M.S. in Bioinformatics at the University of Illinois Urbana-Champaign, M.S. in Biochemistry at Seth G.S. Medical College, and B.S. in Life Sciences at St. Xavier's College.

MICHAEL FITZSIMONS

Michael Fitzsimons, Ph.D. is the Director of Research Programs and Business Development in CTDS.

Michael has a wide range of experience in research and university research administration that he brings to his role at CTDS where he will support and enable research, grant proposals, and publications across the many projects at CTDS. He is an alum of CTDS as he previously served as Director of User Services from 2016-2019.

He earned his Ph.D. in Ecology and Evolution at the University of Chicago and B.S. in Genetics at University of Wisconsin-Madison. His postdoc was in Genome Science at Los Alamos National Laboratory.

FUNDING **SOURCES**

- National Institutes of Health
  - Common Fund
  - National Cancer Institute
  - National Human Genome Research Institute
  - National Heart, Lung, and Blood Institute
  - National Institute of Allergy and Infectious Diseases
  - National Institute of Biomedical Imaging and Bioengineering
  - National Institute of Diabetes and Digestive and Kidney Diseases
  - National Institute of Drug Abuse
  - Office of the Director
  - STRIDES
- Veterans Administration

OTHER NEWS **AND UPDATES**

BLOODPAC FIFTH YEAR ANNIVERSARY

CTDS helped launch the BloodPAC Consortium and led the development of the BloodPAC Data Commons as part of the first Cancer Moonshot in 2016. BloodPAC is celebrating its 5-year anniversary. On March 31, 2022 BLOODPAC hosted a virtual liquid biopsy summit to provide an update on the state of liquid biopsies and outline its strategic plan for the next five years. BloodPAC uses the CTDS-developed Gen3 data platform to host data for its consortium members and for the liquid biopsy community.

GEN3 COMMUNITY FORUM

Over the last three years, we have seen a strong growth in the use of Gen3 to build data commons and data meshes. Today, there are over 30 Gen3 data platforms across the US, Australia and New Zealand. With Gen3 use spreading globally, there is a demand to coalesce shared knowledge and activities into a community. The inaugural Gen3 Community Forum took place over three days, October 10-12, and was co-hosted by the University of Chicago and the Australian BioCommons. This Forum was targeted at individuals and teams who are currently operating a Gen3 data commons or are considering setting one up. In 2023 we plan to hold additional events to further develop the Gen3 community.

GLOBAL ALLIANCE FOR GENOMICS AND HEALTH (GA4GH)

The Global Alliance for Genomics and Health (GA4GH) is working to create frameworks and standards to enable the responsible, voluntary, and secure sharing of genomic and health-related data. CTDS is currently involved with the Cloud workstream with a focus on the Data Repository Service (DRS) API. We are also actively involved in Data Use & Researcher Identities (DURI) workstream discussions related to our contribution to NIH Researcher Auth Services (RAS). CTDS is a member of the steering committee, which votes on all new standards and major updates in all GA4GH workstreams.

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

CTDS PARTICIPATED IN GOOGLE SUMMER OF CODE PROGRAM

Two CTDS projects participated in the 2022 Google Summer of Code Program. GSoC started in 2005 to introduce university students to open source software development. The program has since expanded and connected more than 19,000 new open source contributors from 112 countries with more than 18,000 mentors from 133 countries.

Google Summer of Code

PLANS
FOR THE FUTURE

CTDS is one of the world leaders developing systems and platforms for biomedical data science.

In particular, our open-source Gen3 data platform has been used to build over 20 data commons around the world, including those that CTDS operates, that we operate with our partners, and that third parties operate.

One of our goals for 2023 is to develop technology to dramatically reduce the effort required to bring up and operate a Gen3 data commons so that by the end of 2024 there will be over 120 Gen3 data commons throughout the world.

In 2023, another goal is to develop tools and techniques to build large language models and other deep learning models over the public data in our systems and to use these tools and techniques to make research discoveries.

A third major goal in 2023 is to develop and release a new front end for the GDC and Gen3 so that third party data analysis tools can be easily integrated into these data platforms.

If you are interested in working with us on these or any of other projects, please don't hesitate to reach out.

CONTACT **US**

Visit our website: ctds.uchicago.edu
Explore Gen3 Data Platform: gen3.org
Follow us on Twitter and LinkedIn:

- twitter.com/UChicagoCTDS
- linkedin.com/company/center-for-translational-data-science/

Questions about getting involved or supporting CTDS: ctds@uchicago.edu

THE UNIVERSITY OF CHICAGO
CENTER FOR TRANSLATIONAL DATA SCIENCE