

# CDS User Guide

## Contents

Overview .....	1
How to Submit Data .....	2
Explore Data on the CDS Portal .....	4
CDS Portal Home Page .....	4
CDS Data Page .....	6
Adding Files to the Cart .....	7
CDS Cart Page .....	8
FAQ .....	10

The purpose of this document is to help researchers understand the process of submitting NCI-funded research data to the Cancer Data Service (CDS), as well as accessing and analyzing research data shared by CDS through the CDS Portal.

## Overview

CDS is a data repository within the [Cancer Research Data Commons \(CRDC\)](#) infrastructure for storing cancer research data generated by NCI-funded programs. CDS currently hosts a variety of data types from NCI projects and programs such as the Human Tumor Atlas Network (HTAN), Division of Cancer Control and Population Sciences (DCCPS), and Childhood Cancer Data Initiative (CCDI), as well as data from independent research projects.

CDS provides data-sharing capabilities for additional studies that fall under the following categories:

- Studies with data that do not match any existing CRDC data commons
- Studies with data that do not fit current data type criteria for any CRDC data commons

Studies with data that are on a waiting list to become part of a CRDC [Submission](#) must follow the steps described in “How to Submit Data” on page 2.

Searching the data housed in CDS is easy to do through the Data page, which has a filter function to find data by categories.

Access to controlled data requires approval from the NIH Data Access Committee through the [dbGaP](#) process.

Open-access data is publicly accessible and does not require any approvals. Being a cloud repository, CDS does not facilitate direct downloads of data, owing to high data download charges. However, programs/initiatives can reach out to the [CDS Help Desk](#) to request downloads if they are willing to fund downloads by users approved to access their data. Please contact the CDS helpdesk for more information.

Access to data in CDS for review and/or analysis is through the NCI's Cloud Resource, Velsera Cancer Genomics Cloud (Velsara/CGC) (formerly Seven Bridges). See below for a description of how to select data and import to Velsara/CGC.

## How to Submit Data

The CDS data submission process ensures that CDS is the correct repository to use to share the data and metadata collected and allow basic searches across the data. The process enables data release according to open or controlled access requirements.

### 1. Pre-requisites for data submission:

- a. In the dbGaP registration system, provide the CDS Federal Lead with dbGaP Streamlined access to the study.
- b. Select DCF (Data Commons Framework) as the External Database.
- c. Provide dbGaP with the standardized CDS statement for dbGaP release notes.

*Sample statement: “**Note for data in CDS:** The data for this study are available via the NCI Cancer Data Service. More information about the NCI CDS is available here: <https://datacommons.cancer.gov/repository/cancer-data-service>. The data can be accessed via the Cancer Genomics Cloud (<https://cgc.sbggenomics.com/datasets/file-repository>).”*

- d. Complete the metadata manifest.
  - i. Refer to the [CDE Browser](#) and [caDSR](#) to search for permissible values (PV) of metadata elements.

### 2. Get started:

- a. Complete the CDS New Request Questionnaire.
- b. Contact the CDS Helpdesk email with questions or comments.
- c. When CDS receives the completed New Request Questionnaire, they schedule a data intake interview. It takes one week from the time the questionnaire is received to the time the CDS team schedules an hour-long interview.

### 3. Discuss the following at the initial interview:

- a. Questionnaire responses
- b. Steps taken to de-identify the data for PHI/PII
- c. Requirements for data storage and sharing
- d. CDS Metadata Template

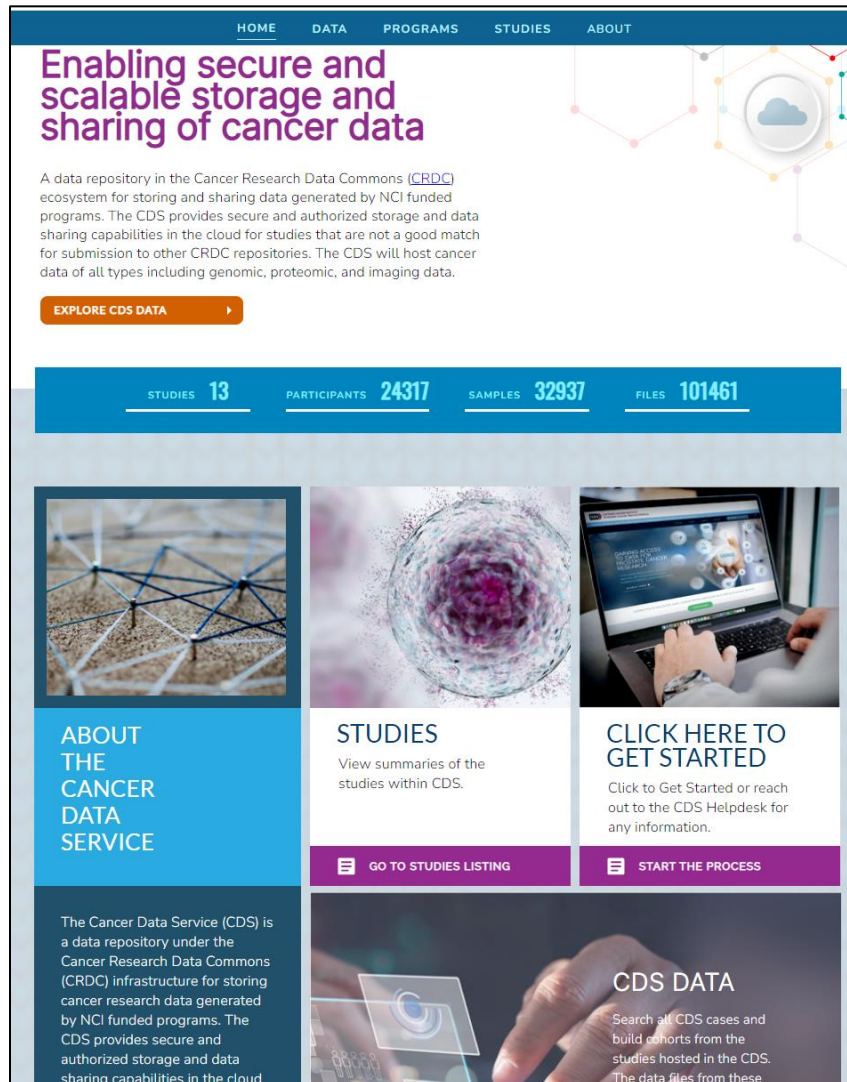
- e. Review pre-requisites, including the dbGaP process, if appropriate.
4. Approval
  - a. CDS notifies the submitter about approval to submit data. This might take a week if new data types or new requirements are included.
5. Submitter submits the following:
  - a. Complete and Final metadata manifest
  - b. List of Authorized Data Uploaders (this includes data uploader name/s, email/s)
6. Once CDS receives the Complete and Final metadata manifest, CDS creates Cloud Buckets.
  - a. CDS sends credentials (for uploader, submission POC, PI). This can take up to one week.
7. The submitter uploads data into the CDS bucket using the provided credentials.
8. The submitter notifies CDS when submission is complete.
  - a. CDS locks the bucket, making it read-only.
9. CDS begins indexing studies.
  - **This process takes 6 to 8 weeks.**
  - The process includes indexing and authentication and authorization (the process of data ingestion for release).
  - CDS releases the study data for secondary sharing on the CDS page on Velsera/CGC (provide link).
  - CDS will try to align the release dates with the dbGaP release schedule as much as possible.
  - The dbGaP study page provides a link to CDS. Data can be explored on the CDS portal and the CDS page on Velsera/CGC.
10. Access the controlled access data on Velsera/CGC, from dbGaP.
  - a. Go to dbGaP and search for the study of interest.
    - On the dbGaP study page you are redirected to the [CDS Data Commons page](#).
    - Note that the data is searchable via Velsera/CGC (formerly known as SBG-CGC).
  - b. Create an account on Velsera/CGC.
  - c. Get approval from dbGaP to access the controlled study.
  - d. Access Cancer Data Service (CDS) File Explore.
  - e. Copy data into your own space within Velsera and start analysis and exploration.
  - f. Visit the CDS page on Velsera/[CGC](#) to see what studies are available and find instructions and guides to use the resources.

## Explore Data on the CDS Portal

Start on the CDS Portal home page to learn how to navigate the services CDS provides.

### CDS Portal Home Page

The following is the CDS Portal home page. A description of its features follows the image.



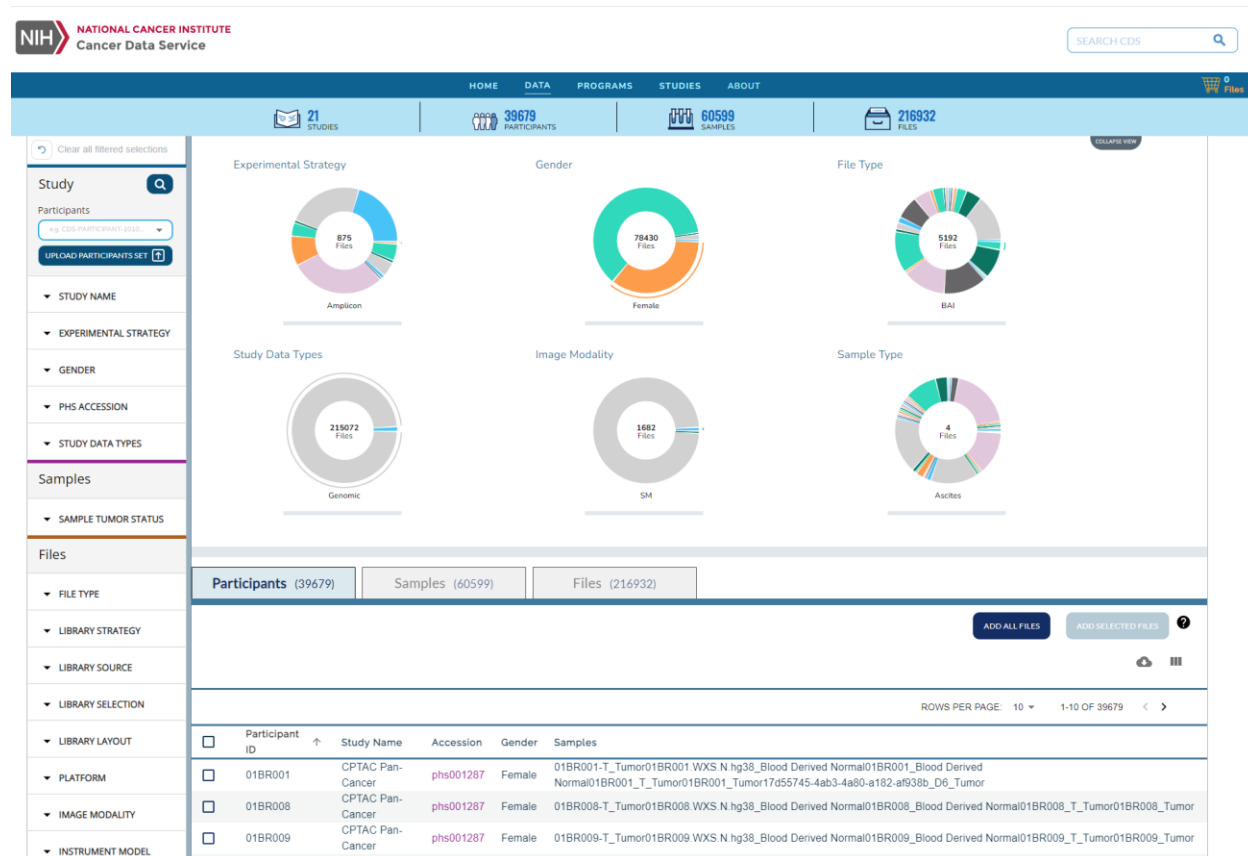
1. **HOME** – Use this page to explore data, programs, studies, and more information about the CDS.
2. **DATA** – This tab opens the Data page where users can search studies using various filters.
3. **PROGRAMS** – This tab opens the Program listing page, where users can learn more about the programs that submit data to CDS, as well as associated studies within the program.

4. **STUDIES** – This tab opens the studies listing page where the user can see the studies that have submitted data to CDS.
5. **ABOUT** – The About tab is a dropdown list that opens several pages with information about CDS, which include:
  - Submission process
  - GraphQL API interface
  - CDS Data Model
  - CDS data and software releases
6. Stats bar
  - **STUDIES** – A count of all the studies in CDS
  - **PARTICIPANTS** – A count of all the participants in CDS
  - **SAMPLES** – A count of all the samples in CDS
7. **FILES** – A count of all the files in CDS

## CDS Data Page

The CDS Data page provides search capabilities across the data housed in CDS. Users can search the data using the various metadata elements associated with the study data.

The following is the CDS Data page. A description of its features follows the image.



1. The following data filters are listed on the left side of the page.

**Note:** CDS is file-centric, so the counts shown for each filter also represent the number of files.

- **STUDY** – The study title that also corresponds with the dbGaP Study Name
- **EXPERIMENTAL STRATEGY** – The type of study or experimental data generated for testing or research purposes
- **SAMPLE TUMOR STATUS** – Normal or Tumor Sample Pathology Indicator
- **GENDER** – This field shows text designations that identify gender. Gender describes a group of people in a society who share particular qualities or ways of behaving which that society associates with being male, female, or another identity.
- **FILE TYPE** – a defined organization or layout representing and structuring data in a computer file

- **PHS ACCESSION** – dbGaP study accession number
  - **STUDY DATA TYPES** – The type of study or experimental data generated for testing or research purposes
  - **LIBRARY STRATEGY** – This field describes the overall strategy for the collection of double-stranded DNA fragments flanked by oligonucleotide sequence adapters to enable their analysis by high-throughput sequencing.
  - **LIBRARY SOURCE** – the source of a sample collection of double-stranded DNA fragments analyzed by high-throughput sequencing
  - **LIBRARY SELECTION** – the type of systematic actions performed to select or enrich DNA fragments used in analysis by high-throughput sequencing.
  - **LIBRARY LAYOUT** – the read strategy or method that was used for sequencing and analysis of a nucleotide library
  - **PLATFORM** – words used to describe the instrument used to carry out a high-throughput sequencing experiment
  - **INSTRUMENT MODEL** – the description of the specific model of the instrument used to carry out a high-throughput sequencing experiment
  - **REFERENCE GENOME ASSEMBLY** – This field stores one or more characters used to identify the published NCBI genetic sequence that is used as a reference against which other sequences are compared.
  - **PRIMARY DIAGNOSIS** – This field is the text term used to describe the patient's histologic diagnosis, as described by the World Health Organization's (WHO) International Classification of Diseases for Oncology (ICD-O).
  - **NUM OF STUDY PARTICIPANTS** – number of participants in the selected study/studies
  - **NUM OF STUDY SAMPLES** – number of samples in the selected study/studies
2. Widgets (the six circles [donuts] on the top half of the screen)
    - **Experimental Strategy** – visual representation of the various types of experimental strategies in the CDS Portal
    - **Gender** – visual representation of the files for various genders in the portal
    - **File Type** – visual representation of all the count of file types in CDS
    - **Study Data Types** visual representation of the count of all the study data types in CDS
  3. Data Table (the light blue header across the top of the screen)
    - **Participants** – number of participants in the study/studies
    - **Samples** – number of samples in the study/studies
    - **Files** – counts of the files

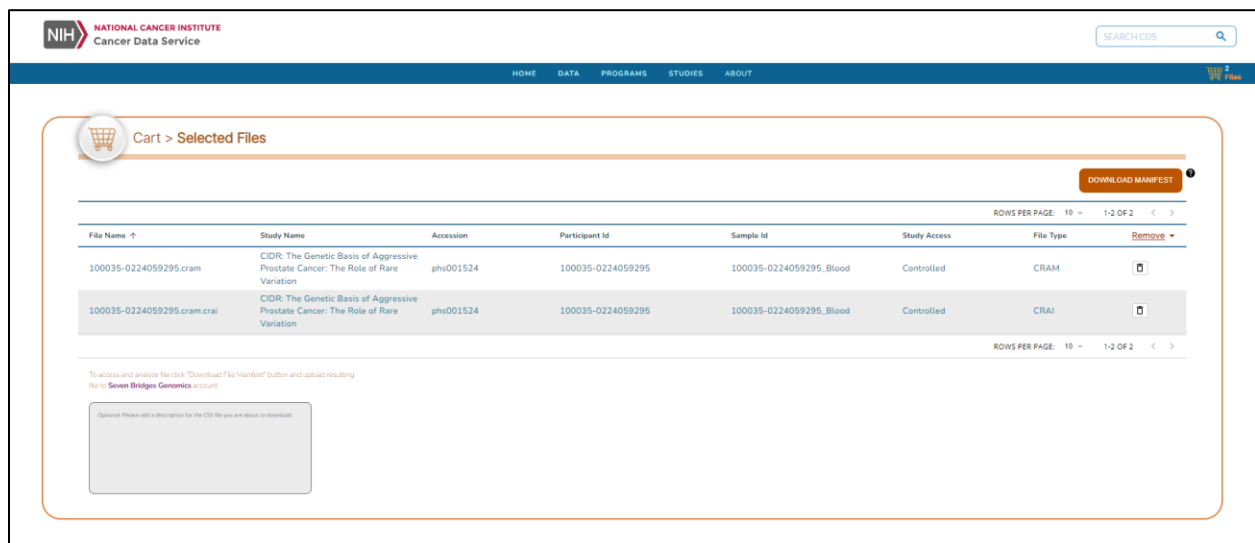
### [Adding Files to the Cart](#)

Files of interest for analysis can be selected and added to the cart as follows:

1. Select the study or PHS of interest.
2. Select all filters/metadata elements of interest.
3. The selected files appear in the table below the cart (lower-right section of the screen).
4. Click the box for the files to be added to the cart.
5. Click **ADD SELECTED FILES**, which loads the selected files to the cart.
6. Click the cart icon on the top right of the page, which opens the cart page.

## CDS Cart Page

Using the Cart page, users can download a data manifest to import to the Velsera/SBG-GCG cloud resource.



The screenshot shows the NIH Cancer Data Service Cart page. At the top, there is a navigation bar with links for HOME, DATA, PROGRAMS, STUDIES, and ABOUT. A search bar is located on the right. The main content area is titled 'Cart > Selected Files' and features a 'DOWNLOAD MANIFEST' button in the top right corner. Below this is a table with the following columns: File Name, Study Name, Accession, Participant Id, Sample Id, Study Access, and File Type. The table contains two rows of data. Below the table, there is a text box for a description and a 'Download Manifest' button.

File Name	Study Name	Accession	Participant Id	Sample Id	Study Access	File Type
100035-0224059295.cram	CIDR: The Genetic Basis of Aggressive Prostate Cancer: The Role of Rare Variation	phs001524	100035-0224059295	100035-0224059295_Blood	Controlled	CRAM
100035-0224059295.crai	CIDR: The Genetic Basis of Aggressive Prostate Cancer: The Role of Rare Variation	phs001524	100035-0224059295	100035-0224059295_Blood	Controlled	CRAI

1. While on the Cart page, create a description for your file in the text box on the bottom left of the screen (this is optional).
2. Click the **DOWNLOAD MANIFEST** button in the top-right corner of the Cart page.
3. Open the downloaded folder on your computer. The manifest will contain the following metadata elements:
  - DRS URI (File ID)
  - Name
  - Participant ID
  - Md5sum
  - User Comments

Import the manifest for analysis on Velsera/CGC as follows:

1. Create an account or log in at the following URL:  
<https://www.cancergenomicscloud.org/>.
2. Create a new project or select an existing project suitable for digital access to the files listed in the file manifest.



3. Navigate to the CGC dashboard's navigation bar, choose **Files**, and click the **+ Add files** button in the dropdown menu.
4. From the dropdown menu, select **GA4GH Data Repository Service (DRS)**.
5. Choose **From a manifest file** and click **Browse manifest** to import the files with the DRS URI attached, generated from the CDS portal. (You can typically find this file in your downloads folder or on the desktop for easy access.)
6. Use the search box to add relevant tags or comments associated with the batch of imported files and click **Import**.
7. Check the Data Use Agreement box to confirm compliance with the data guidelines set by the CGC. Then, click **submit** to access your files in the created project.
8. Use the search box to create tags to associate with your data.  
DRS (Data Repository Service) identifiers are now displayed for each file from the imported file manifest within the CGC project you selected.  
You can access these files in the CGC Genome Browser and can select them as inputs for downstream analysis.

### CDS Cart Page

Users can access the GraphQL interface using Jupyter or by directly querying the interface. Below are a few queries which can be run directly from the GraphQL interface on CDS:

#### Query #1: Retrieve all information in CDS

```
{
  file{
    file_id,
    file_name,
    file_description
  }
}
```

#### Query #2: Finding all files in a specific study:

```
{
  study(phs_accession: "phs002371") {
    files{
      file_id,
      file_name,
      file_description
    }
  }
}
```

### Query #3: Retrieving files associated with a specific patient:

```
{
  participant(participant_id:"PBBJUH"){
    dbGaP_subject_id,
    samples{
      sample_id,
      files{
        file_id,
        file_name,
        file_description
      }
    }
  }
}
```

Prior to submitting the data, you must submit a manifest of your submission's metadata. This is a summary of what is in your submission and includes data such as filenames, MD5 checksums for the files, and file size. CDS uses the metadata manifest to index the data so it can be made available via NCI Cloud Resources. Once CDS curators have approved the metadata manifest, you will receive credentials to start your data submission in your bucket on Cloud One. Please note that the indexing process is time-consuming and can take anywhere from six to eight weeks.

Once you are ready to start your submission, CDS curators will send you instructions on how to upload your data to a cloud S3 bucket. The CDS team conducts open office hours and can walk you through the technical details to help you get started. Once you are done uploading, reach out to the CDS team to let us know and we will lock the bucket (you will continue to have read-only access to the bucket).

After the indexing is complete, the data will be available through NCI's Cancer Research Data Commons (CRDC). NCI requires authentication through [eRA Commons](#) and [dbGaP](#) authorization to access controlled data.

## FAQ

### Q: How is the data made available?

- Through NCI Cloud resources: access analysis tools, workflows, and workspaces
- Through NCBI (dbGaP/SRA)
- Both controlled-access and open-access data
- Access is controlled based on dbGaP allow lists.
- General rule: direct data download is not supported.

### Q: What data is eligible for CDS?

- Data from NCI-funded programs, which is being made available for secondary data sharing
- Accepts various data types (e.g., genomic, imaging)

**Q: How long does the submission process take?**

- It can take up to one week to create a bucket and credentials for a submission.
- Indexing: After you send us the metadata manifest and the uploading is complete, it takes about 6 to 8 weeks to release the data.