CCDI Data Ecosystem: Data & Tools Submission Guidance

Introduction

The Childhood Cancer Data Initiative (CCDI) focuses on the critical need to collect, analyze, and share data to address the burden of cancer in children, adolescents, and young adults (AYAs). The initiative supports maximizing the use and benefit of data and tools from childhood and AYA cancer research to improve outcomes for patients and survivors. In line with NIH policies, the CCDI expects submitters to share data and tools in a consistent and compliant manner.

Purpose

This document provides guidance for prospective submitters on how to contribute data and tools to the CCDI Data Ecosystem. It outlines the submission process, key pre-requisites, best practices, and resources to help ensure a smooth and successful submission experience.

For questions, please contact at: NCIChildhoodCancerDataInitiative@mail.nih.gov

Expectations

Sharing childhood cancer data with the broad scientific community is a central goal of the CCDI. Data and tools must be prepared to ensure broad usability and are required to comply with NIH Data Management and Sharing policy and CCDI programs expectations. To that end, CCDI expects that data, metadata, algorithms, tools, code, and other relevant resources will be shared with the wider scientific community in a timely and accessible manner.

Submitters will be required to abide by the CCDI program expectations, which includes:

- Human subject data must be de-identified in accordance with legal and ethical standards, including the HHS HIPAA De-identification Guidelines.
- Data must comply with the NIH Data Management and Sharing Policy and other applicable data sharing policies.
  - Data must be accessed, manipulated and analyzed by end users using open-source codes or tools without license restrictions and fees, to the greatest extent possible.
  - Disclosure of any applicable factors affecting subsequent access, distribution, or reuse of scientific data related to informed consent (e.g., disease-specific limitations, particular communities' concerns), and privacy and confidentiality protections (i.e., de-identification, Certificates of Confidentiality, and other protective measures) consistent with applicable federal, Tribal, state, and local laws, regulations, and policies.

- Any data analysis tools or data processing tools developed are expected to be open-source and either deposited into GitHub or incorporated into the NCI Cloud Resources, e.g. Cancer Genomic Cloud, etc.

Submitting Data to the CCDI Data Ecosystem

Studies submitted to CCDI are indexed in the CCDI Hub Explore Dashboard at the file level. The dashboard displays row-level metadata for participants, diagnoses, studies, samples, and files, which users can explore and use to build cohorts.

For CCDI, some studies contain open-access data, while others contain **controlled-access datasets** (i.e., datasets containing identifiable human data). For human data, access is governed by terms and conditions consistent with the participants' informed consent. Confidentiality and participant privacy are always maintained.

The process for submitting data to the CCDI Data Ecosystem generally follows the steps outlined below:

1. Completion of the Metadata Submission Template

Submitters must complete the **CCDI Submission Template**, an Excel workbook containing sheets corresponding to nodes in the CCDI Data Model. The required sheets vary depending on the data types being submitted (e.g., sequencing files, radiology files).

Required sheets for every study include:

- Study
- Study_admin
- Study_funding
- Study_personnel
- Publication
- Participant
- Diagnosis

Any additional applicable participant sheets, such as:

- treatment
- treatment_response
- survival

Any applicable data file node sheets, such as:

- radiology_file
- sequencing_file

- methylation_array_file
- cytogenomic_file
- clinical_measure_file
- pathology_file
- generic_file

All required fields (highlighted in yellow in Metadata Submission template) must be completed to ensure accurate and timely submission. Submitters should refer to the README and INSTRUCTIONS sheet in the workbook for detailed guidance, including summaries of changes by version. The workbook also includes reference sheets for the CCDI Data Dictionary, Term Definitions, and Value Sets.

Note for Imaging Data Submissions: For open-access imaging data (e.g., H&E slides) that will be uploaded to NCI's designated Google Cloud Storage buckets, submitters must also complete the Imaging Data Risk Mitigation Considerations form.

2. Iterative Metadata Review

CCDI Data Curation team work collaboratively with submitters in an iterative process to review and finalize the metadata. This ensures completeness and alignment with CCDI data standards prior to ingestion into the ecosystem.

Controlled-Access Data Submission and dbGaP Registration

For studies that include controlled-access data (e.g., potentially identifiable human data), the following steps apply:

- Studies will be registered in dbGaP (Database of Genotypes and Phenotypes).
- Submitters must complete the CCDI Metadata Template, along with the Study Data Outline, Study Configuration File, and Institutional Certification.
- CCDI Data Curators will generate the required dbGaP files, including Subject and Sample Attributes and Subject-Sample Mapping files, extracted from the metadata template.
- The NCI Genomic Program Administrator (GPA) manages the study registration process in dbGaP.
- Submitters upload study data files (e.g., to AWS S3 buckets) for indexing in the CCDI Hub Explore Dashboard.

A general workflow for controlled-access study data submission and release is as follows:

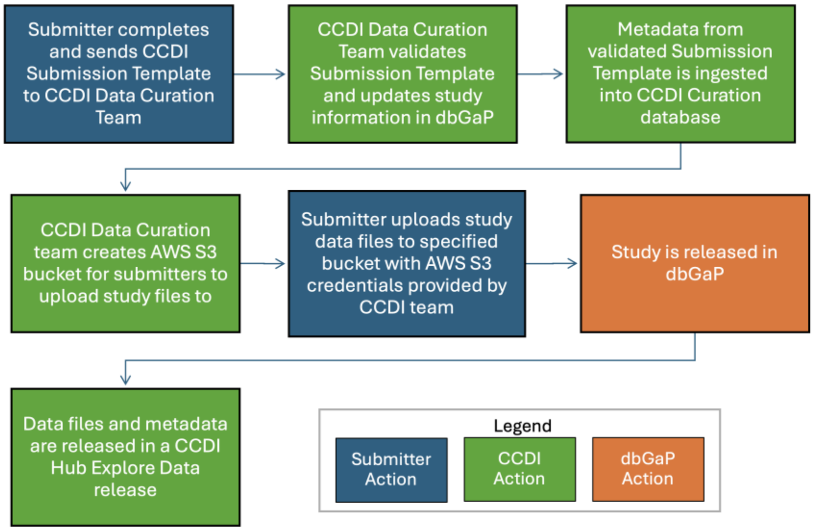

Submitter completes and sends CCDI Submission Template to CCDI Data Curation Team
CCDI Data Curation Team validates Submission Template and updates study information in dbGaP
Metadata from validated Submission Template is ingested into CCDI Curation database
CCDI Data Curation team creates AWS S3 bucket for submitters to upload study files to
Submitter uploads study data files to specified bucket with AWS S3 credentials provided by CCDI team
Study is released in dbGaP
Data files and metadata are released in a CCDI Hub Explore Data release
Legend
Submitter Action
CCDI Action
dbGaP Action

Figure 1: General workflow for controlled-access study submissions

Please note that open-access studies will not require dbGaP study submission or dependence on study release in dbGaP for release in CCDI.

The CCDI Data Curation team will validate all submission components and provide support throughout the process. For projects updating an existing dbGaP study (e.g., adding new data types or participants), please contact us to coordinate next steps.

Data Best Practices

Adhering to data management best practices ensures that submitted data adhere to FAIR principles (findable, accessible, interoperable and reusable). This enhances the scientific value and impact of the data by facilitating data reuse and integration across pediatric cancer studies. Well-curated data accelerate discovery, foster collaboration, and support meaningful outcomes for pediatric patients.

We encourage adherence to the following Data Best Practices:

- **Use Clear and Consistent File Naming Conventions**
  - Include meaningful components such as subject ID, data type and file extension/file format.
  - Example: PedsOnc_1234_20240601_RNAseq.fastq
  - For de-identified clinical reports and other files containing a single participant’s data, include the study participant identifier or global identifier, e.g. COG Universal Subject Identifier (USI), in the file name.
- **Use/Submit Data from Standard File Formats**

    - Submit data in CCDI acceptable formats, please see the list of accepted file formats in the CCDI Data Model.
    - Avoid proprietary or outdated formats.
- **Include Comprehensive Metadata**
    - Provide detailed and structured metadata, including clinical variables, sample characteristics, processing methods, and study protocols.
    - Align with community standards such as NCI Cancer Data Standards Registry (caDSR) and use CCDI common data elements (CDE) wherever possible.

- **De-identify Data Thoroughly**
    - Remove all direct identifiers in accordance with the HIPAA Safe Harbor Method.
- **Follow Controlled Vocabularies and Ontologies**
    - Use standardized terminologies (e.g., ICD-O-3, WHO CNS5) to describe diagnoses, procedures, and phenotypes.
    - This ensures data interoperability with NCI repositories and other datasets.
- **Enriched Sequencing File Library Metadata**
    - For sequencing data files, each library_id value in the Metadata Submission template must be unique, and the value should allow users to identify and distinguish your sequences.
    - The design_description property should be treated as a materials and methods section describing how the library was prepared and sequenced (i.e. 1 to 3 sentences without newlines or special characters).

- **Validate and QC Data Before Submission**
    - Perform quality checks and validation to ensure completeness, accuracy, and consistency. Tools like FastQC (for sequencing) and schema validators can be helpful.

File Upload to AWS S3 Buckets

For data files (e.g. BAM, CRAM, VCF, etc.), clinical report files and other files that are intended to be shared and findable in the CCDI Hub Explore Dashboard portal, these are typically submitted to CCDI via upload to a designated AWS S3 bucket. The CCDI Data Curation team will provide submitters a set of credentials to perform upload to designated AWS S3 bucket(s) for data upload. Note: To protect data and data provenance, once upload of study data has been completed, submitter AWS credentials will be converted to read-only access.

Prior to uploading files to bucket, it is recommended that file organization and folder/directory structure are organized in a meaningful and systemized manner to reflect files belonging to same participant, study arm, molecular assay etc. Note: There are no restrictions on file sizes or number of files for upload to AWS S3 buckets.

Please see the Appendix: AWS S3 File Upload User Guide for directions on how to upload study files to AWS S3 buckets.

Submission and Additional Resources

Data Sharing Policies and Best Practices Resources

HHS HIPAA De-identification Guidelines

NCI Data Sharing Policy Guidance

NIH Data Sharing Policies

CCDI Resources

CCDI Hub

CCDI Data Model

CCDI Submission Template Location in CCDI Data Model Repo

CCDI Preferred Data Elements

C3DC Data Model – Data model for Childhood Cancer Clinical Data Commons (C3DC), a repository of harmonized clinical data from CCDI and other pediatric cancer data sets.

CCDI Contact Email: NCIChildhoodCancerDataInitiative@mail.nih.gov

Appendix: AWS S3 File Upload User Guide

Prerequisites

1. The csv-formatted file sent by CCDI data management team containing your credentials. The important fields in the file are:
    a. the “**Access key ID**”
    b. the “**Secret access key**”
2. The **name of the S3 buckets** provided to you by CCDI
3. AWS CLI Installed on your computer (details below)

AWS CLI Installation

AWS Command Line Interface (AWS CLI) is a set of tools that allow you to interact with AWS resources (including S3 buckets) from the command line (e.g. terminal/command-prompt). More information from AWS here: https://aws.amazon.com/cli/

To install AWS CLI, please see Amazon’s documentation in the following link: https://docs.aws.amazon.com/cli/latest/userguide/getting-started-install.html

Also see the following for information on installing AWS CLI: https://github.com/aws/aws-cli/tree/v2#installation

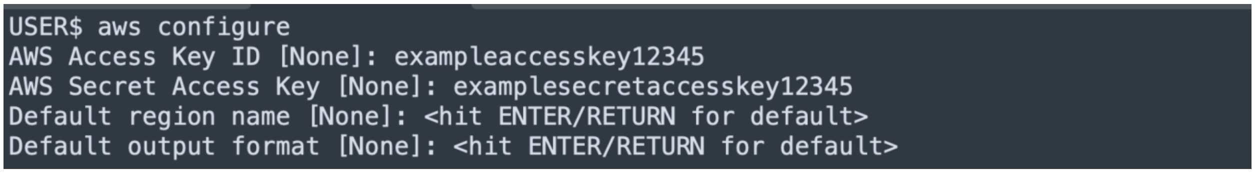
Setting up AWS Profile (Loading Credentials)

Once AWS CLI is installed, the next step is to set up the AWS CLI profile on your computer. You only need to configure your profile once. This process will store the credentials for accessing your S3 buckets on your computer.

To configure your profile, run the following command in your terminal:

```
aws configure
```

- The command will ask for “AWS Access Key ID”, use the “Access key ID” from the credentials file.
- The command will ask for “AWS Secret Access Key”, use the “Secret access key” from the credentials file.
- For default region name, you can just use the default value (just hit enter to use the default value).
- For default output format, it is recommended to just use the default value (just hit enter to use the default value).



```
USER$ aws configure
AWS Access Key ID [None]: exampleaccesskey12345
AWS Secret Access Key [None]: examplesecretaccesskey12345
Default region name [None]: <hit ENTER/RETURN for default>
Default output format [None]: <hit ENTER/RETURN for default>
```

Figure 2: An example terminal session configuring AWS profile (aws configure)

Figure: Verify Access

To verify that you can access the AWS S3 buckets from the command line, you can use the "aws s3 ls" command to list the contents of the buckets. This command will list the contents of the bucket (e.g. think "ls" command). This command will work even if the bucket is empty (will return 0 files).

You will need the name of the bucket you are trying to access. The bucket *may* have a name like "cds-123-phs004567" or something similar.

To test access, run the following command:

```
aws s3 ls s3://<bucket_name>
```

If this command does not return any error, it means you have access to the bucket.

Uploading (copying) Data to Buckets

To copy a local file to S3 bucket, use the "aws s3 cp" command. This command copies a single file to a specified bucket.

For example, if I wanted to upload a file called "test.txt" to the bucket named "cds-123-phs004567":

```
aws s3 cp test.txt s3://cds-123-phs004567/test.txt
```

Expected Output:

```
upload: test.txt to s3://cds-123-phs004567/test.txt
```

Note: this command can also be used to rename the file, just like the Unix "cp" command.

To upload an entire directory to an S3 bucket, you can use the "--recursive" option, for example:

```
aws s3 cp ./some-directory s3://cds-123-phs004567/some-directory –-recursive
```

AWS will then create the directory "some-directory" and copy all of the local contents of the "./some-directory" into it.

You can view files you have uploaded so far to the bucket and their full paths/directory structure of the uploads be performing the following command:

```
aws s3 ls s3://cds-123-phs004567/–-recursive
```

More Examples

https://docs.aws.amazon.com/cli/latest/reference/s3/cp.html

https://docs.aws.amazon.com/cli/latest/userguide/cli-services-s3-commands.html