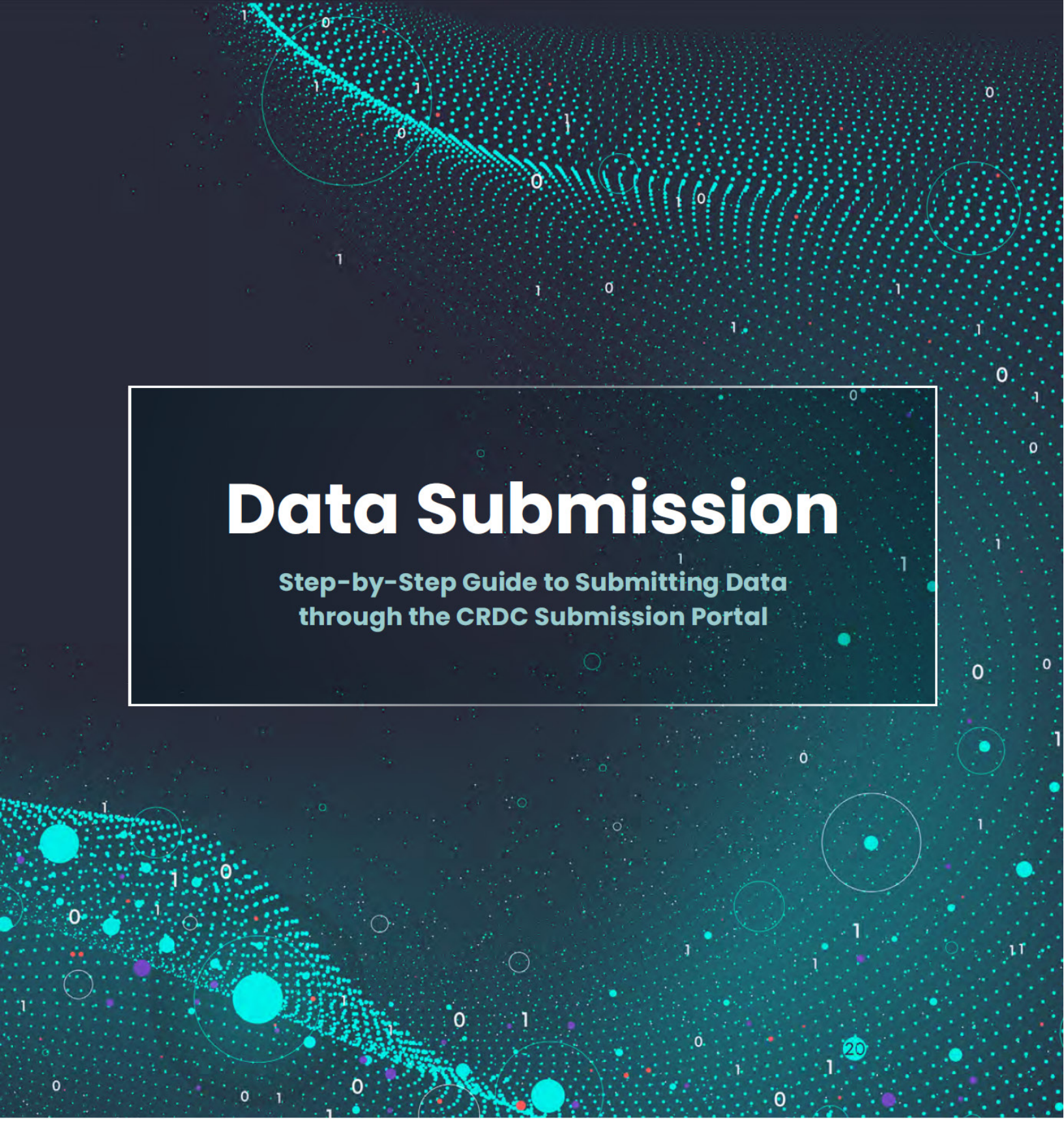

Data Submission

Step-by-Step Guide to Submitting Data through the CRDC Submission Portal

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

datacommons.cancer.gov

Table of Contents

I. INTRODUCTION 1

II. PREREQUISITES 1

III. CONDITIONAL APPROVAL 1

IV. ACCESSING DBGAP SHEETS IN THE SUBMISSION PORTAL 2

V. CRDC DATA MODELS 3

VI. DOCUMENTATION 3

VII. REQUEST ACCESS 4

VIII. STARTING A NEW SUBMISSION 5

IX. CONTINUING AN EXISTING SUBMISSION 8

1. Adding and Managing Collaborators from the Dashboard 8
2. Obtaining Submission Templates 9
3. Downloading Data Dictionary and Submission Templates 12
   - 3.1 Submission Templates and Properties 14
     - 3.1.1 Special Columns 15
     - 3.1.2 Type Column 15
     - 3.1.3 Relationship Columns (Parent Mapping Columns) 15
4. Uploading Data Files and Metadata Manifests 15
   - 4.1 Uploader CLI Tool 15
     - 4.1.1 Introduction 15
   - 4.2 Downloading the Uploader CLI Tool 16
     - 4.2.1 Download the Uploader CLI Tool from the CRDC Submission Portal 16
     - 4.2.2 Cloning the Uploader CLI Tool from GitHub 17
   - 4.3 Setting Up the Python Environment for CLI (Source version) 17
   - 4.4 Using the Uploader CLI Tool 18
     - 4.4.1 Configuration File in YAML Format 18
     - 4.4.2. File Manifest in TSV Format 21
     - 4.4.3. Archive Manifest in TSV Format 22
   - 4.5 Starting the Upload Process 22
   - 4.6 Using the CRDC Submission Portal's Graphical Interface to Upload Metadata Submission Templates 23
5. Running Validations 25
   - 5.1 Reviewing Validation Results 26
     - 4.1.1. Viewing and Filtering Validation Results 26
     - 4.1.2. Requesting Permissible Values (PV) 28

5.2 Revising Released Metadata – Options to Keep or Overwrite Existing Content 30
5.3 Correcting Errors 31
5.4 Remove Specific Files 32
6. Submitting Your Final Dataset 32
7. What to Expect After Submission 33

I. INTRODUCTION

This tutorial walks you through the process of submitting data to CRDC through the CRDC Submission Portal. If you have questions that are not answered here, email the CRDC Help Desk (NCICRDC@mail.nih.gov) or contact the Data Concierge assigned to your submission once you successfully create a submission.

II. PREREQUISITES

Before starting your data submission, complete the following prerequisites:

- **Secure approval from the CRDC Submission Review Committee to submit data.** Notification of approval or rejection appears on the CRDC Submission Portal under Submission Request and in an email sent to the requestor. If rejected, consider other repositories for sharing data at NIH.

- **Create a Login.gov account.** It is strongly recommended that the Login.gov identity be associated with the submitter's/user's organization or institution; however, it is not a requirement. Using an institutional email as your user identity helps us quickly determine your organization, but you may choose a personal email instead. NIH staff can log in using their PIV card.

  Note: If you do not log in to the CRDC Submission Portal account within 60 days, your access to the portal will be deactivated. This applies even while you are working on an active, ongoing submission. To reactivate your access, contact the CRDC Help Desk (NCICRDC@mail.nih.gov).

- **Request Access to the Submitter role.** Users must request the *Submitter* role to submit data to the CRDC Submission Portal or oversee the relevant submissions. See "VII. REQUEST ACCESS" on page 4 for more details.

The CRDC Submission Portal uses CRDC standard Common Data Elements (CDEs), and all submissions are expected to use these CDEs and comply with their permissible values. A comprehensive list of CRDC standard CDEs can be found at caDSR. On the caDSR website, click the **CRDC Standard Data Elements** link in the **Links to Favorites** section or download them using the *getCRDCList* endpoint of the caDSR API. It is recommended that submitters familiarize themselves with CRDC standards before starting a submission.

III. CONDITIONAL APPROVAL

To submit data to CRDC, users are required to first complete and submit the submission request form for approval. The form is reviewed by the Submission Review Committee (SRC). If the Request Form is conditionally approved, the submitter must provide the following required information, if applicable, by emailing the CRDC Help Desk (NCICRDC@mail.nih.gov). The submitter can initiate the data submission only after all the conditions have been resolved.

Required Information, if applicable:

- **dbGaP ID/Accession number (phs00####):** If the submission contains controlled access data, the study must be registered at the database of Genotypes and Phenotypes (dbGaP). Upon registration, dbGaP will assign a dbGaP ID/Accession number (phs000####) to the study. The CRDC Submission Portal will not allow users to submit controlled access data without a dbGaP ID/Accession number. **Note:** Data will be released on the respective CRDC Data Commons portal after the study is released on dbGaP. It is therefore recommended to register the study and work on dbGaP submissions concurrently with the submission to CRDC.
  - **Name of the Genomic Program Administrator (GPA):** Providing the name of the GPA who registered the study in dbGaP is required before initiating the data submission in the CRDC Submission Portal.
  - **Data Model Update**: In case the study requires data model changes due to the introduction of new data types, the CRDC Data Commons team must implement necessary updates before submissions can proceed. In such cases, the team will request clarification from the listed Principal Investigator (PI) on the conditionally approved Submission Request Form. Once the data model update is complete and the condition is cleared, the PI will be notified via a system-generated email.

IV. ACCESSING DBGAP SHEETS IN THE SUBMISSION PORTAL

To help the submitters working on submissions to CRDC and dbGaP simultaneously, the CRDC Submission Portal generates partially populated dbGaP sheets based on the data uploaded to the portal. Note that these documents are not final and are intended to be a preliminary aid for your data submission to dbGaP. Submitters are responsible for verifying and completing all the necessary fields required for a dbGaP submission.

The portal will automatically make the **dbGaP sheets** available for download once the first batch of metadata files are uploaded to the portal. See Figure 1. Although the dbGaP sheets are available throughout the data upload process, downloading them after **validations are complete** on the CRDC portal is strongly recommended as validations offer useful feedback for verifying data accuracy.

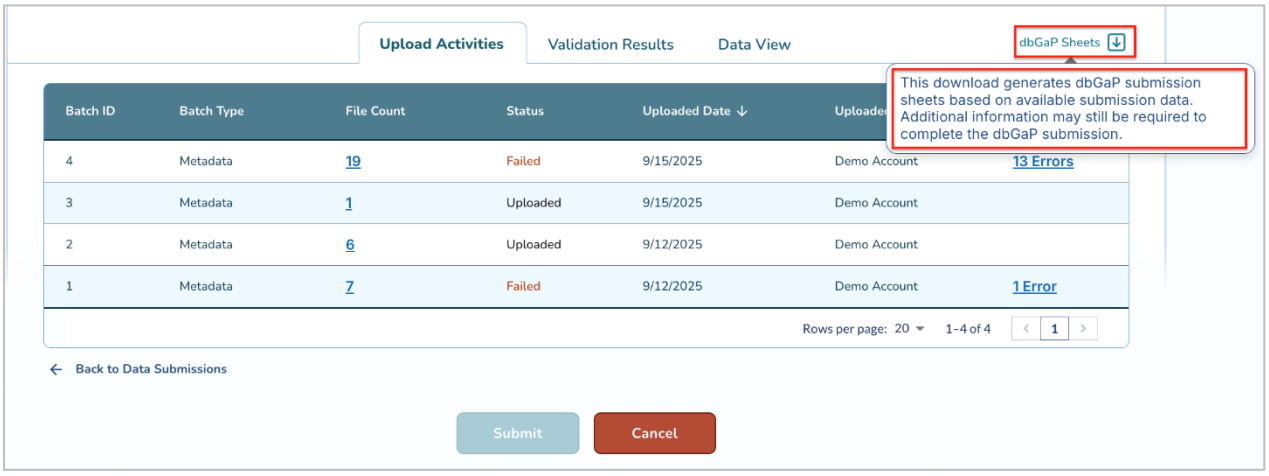

Figure 1. Download dbGaP Sheets

V. CRDC DATA MODELS

CRDC's Data Commons use data models to organize data in a consistent and structured manner, ensuring accuracy and facilitating reusability. These data models are graph-based, and data are organized as nodes, properties, and relationships. The data models supported by CRDC are displayed in the Data Model Viewer of the CRDC Submission Portal; each Data Commons (DC) has its own specific data model. See Figure 2.

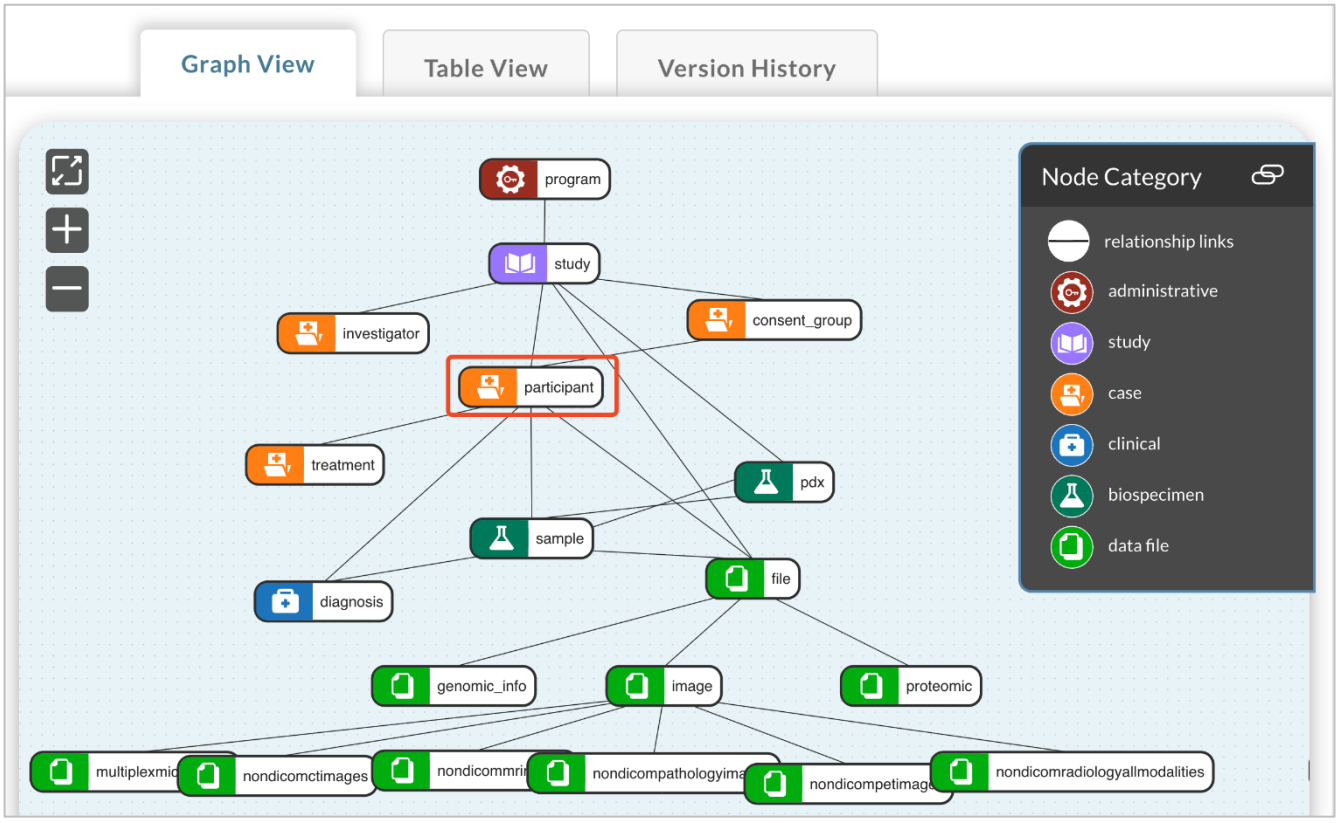

Figure 2. Participant Node Relationships in the GC Data Model

VI. DOCUMENTATION

Submitters can find this document, as well as instructions on using APIs to submit data, under the **Documentation** tab of the CRDC Submission Portal.

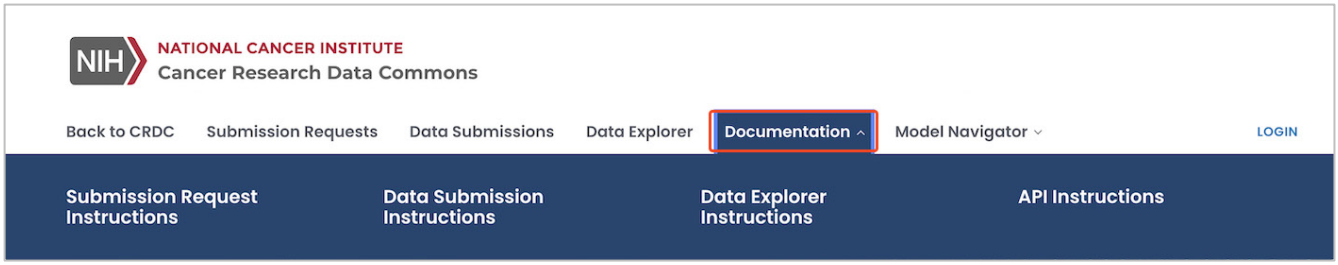

Figure 3. Documentation Menu Showing Available Documents

VII. REQUEST ACCESS

1. The Request Access button can be used to make the following requests:
   a. Submitter role (required to submit data)
   b. Access to specific studies, or
   c. Update the Institution Name associated with the CRDC Submission Portal account.
2. Log in to the CRDC Submission Portal and go to your user profile by clicking your name in the upper-right corner of the page.
3. By default, each account is assigned a **User** role. To begin the data submission, the user must request the **Submitter** role by clicking the **Request Access** button. See Figure 4. A **Request Access Form** appears, as shown in Figure 5, where you can provide the required information.

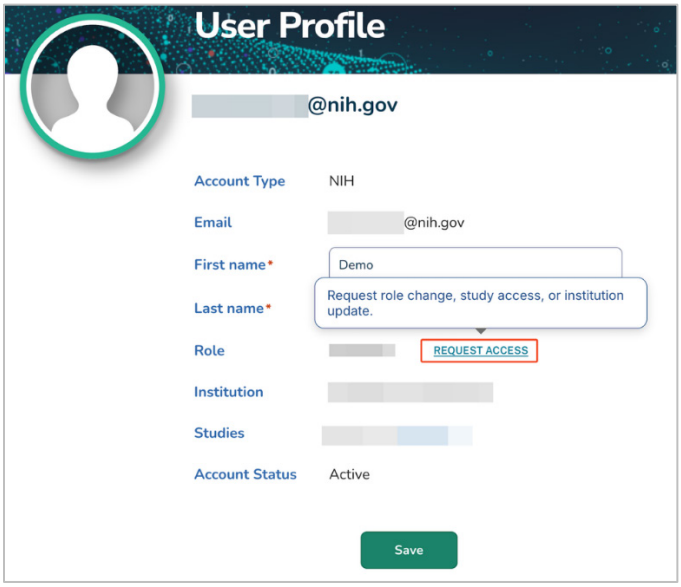

Figure 4. Request Access

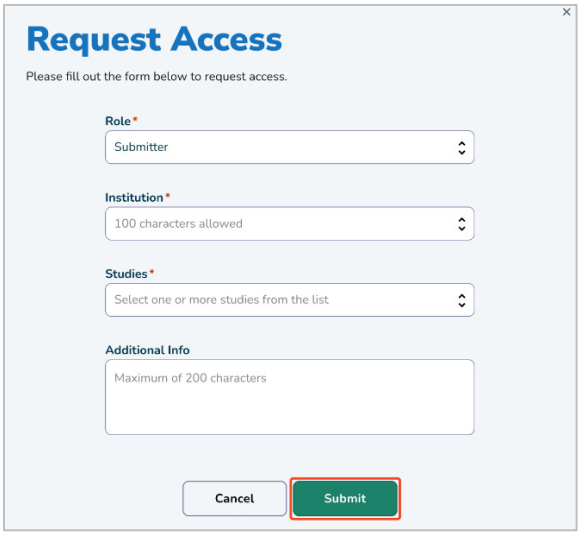

Figure 5. Request Access Form

4. From the **Role** dropdown menu, select **Submitter** if you want to work on the data submission or oversee the submissions for specific studies.
5. From the **Institution** dropdown menu, select your institution's name. If not listed, enter it manually. In the **Studies** dropdown menu, select the relevant studies from the list.
6. Optionally, users can provide additional details about their role in the **Additional Info** box before submitting the form.

   The CRDC System Administrator accepts the request, and the user is notified via their login identity. The user who was granted the **Submitter** role can now start the data submission process.

Note: If you do not log in with the CRDC Submission Portal account within 60 days, your access to the portal will be deactivated. If that happens, contact the CRDC Help Desk (NCICRDC@mail.nih.gov) to reactivate your account.

VIII. STARTING A NEW SUBMISSION

Once logged in with a Submitter role, navigate to the Data Submissions tab on the CRDC Data Submission portal. The submitter is taken to the Data Submission List page. If this is the Submitter's first data submission, the table showing the list of Data Submissions will be empty. If the Submitter has multiple submissions, use the filters at the top of the table to narrow down the list. See Figure 6.

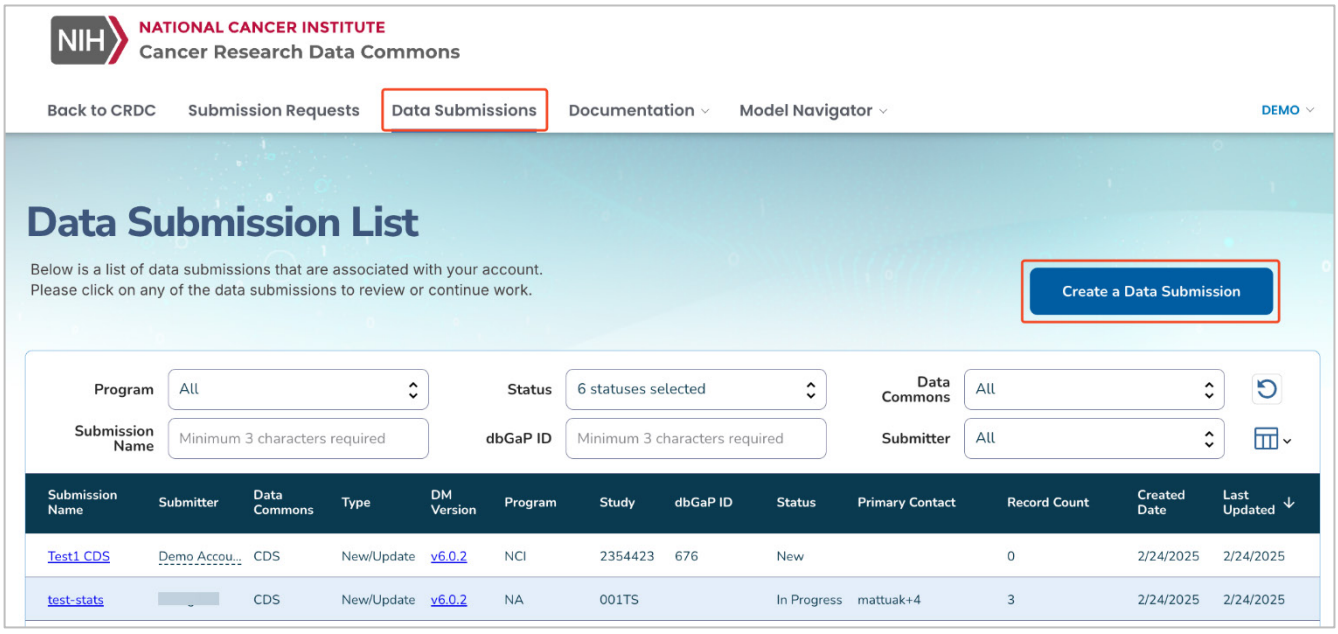

Figure 6. Create a Data Submission

To start a new data submission, click the **Create a Data Submission** button. A dialog box as shown in Figure 7 appears. Fill out all the required information as described below.

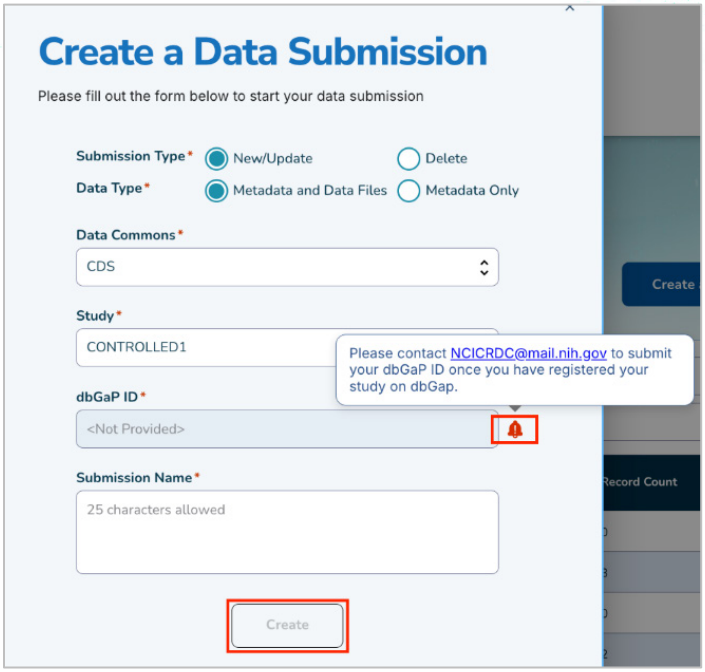

Figure 7. Data Submission Dialog Box

1. Choose the Submission Type.
   - Select **New/Update** to create a new data submission or update an existing one.
   - Select the **Delete** option to remove files from a previous submission already released publicly by the Data Commons. Selecting the Delete option keeps only one Data Type option enabled, which would be **Metadata Only.** Submit the metadata associated with the data that needs to be deleted. The deletion request goes through a validation process by the CRDC team and the CRDC Data Commons. Once approved by the CRDC Data Commons, the deletion is processed.
2. For Data Type, indicate whether you are submitting both metadata and data files or only metadata files by selecting one of the options, **Metadata and Data Files** or **Metadata Only**, respectively.
3. Select the **Data Commons** that you were approved to submit to by the Submission Review Committee (SRC) for your data, if it is not already preselected. If you do not see the CRDC Data Commons listed, email the CRDC Help Desk (NCICRDC@mail.nih.gov).
4. The **Study** dropdown menu displays the Study Title (or the Study Abbreviation) you previously shared through the Submission Request Form. If you notice an error in this list, email the CRDC Help Desk (NCICRDC@mail.nih.gov).
5. If your study includes controlled access data, the **dbGaP ID** /accession number will be pre-populated as provided on the Submission Request Form. If you did not provide the dbGaP ID on the Submission Request Form, please email it to the CRDC Help Desk. The system will not allow the submitter to initiate data submission if the submitter has not shared the dbGaP ID.

   Additionally, if your submission request was conditionally approved due to a missing GPA name in the Submission Request Form or because your study requires a data model change, the **Create a Data Submission** button will remain disabled until the condition is cleared by the system administrator.

6. Submitters can give the submission a **Name** in the provided free-text field to label their submissions. This name appears in the Submissions List table on the Data Submissions List page, once the submission is created. A submission name which is relevant and informational is highly recommended.

Click the **Create** button to create the new submission. This button will only be enabled if the conditions for submission are cleared. Once a submission is created, CRDC assigns a Data Concierge to your submission. You can find the email address of the Data Concierge assigned to your submission on the dashboard of your data submissions page. **From this step onwards, all questions related to the data submission should be directed to the assigned Data Concierge.**

IX. CONTINUING AN EXISTING SUBMISSION

To access and update an existing submission, go to the **Data Submissions** tab. A table, listing all the existing submissions, appears on that page. Under the **Submission Name** column, select the submission you want to continue with. To see the full version of the submission name, hover over the name (see Figure 8). The submitters can customize the columns displayed in the submission table. Desired columns can be selected and added by clicking the **Table** icon in the top-right corner and applying the changes. The filters at the top of the table provide useful ways to refine searches, especially when the list of submissions is extensive.

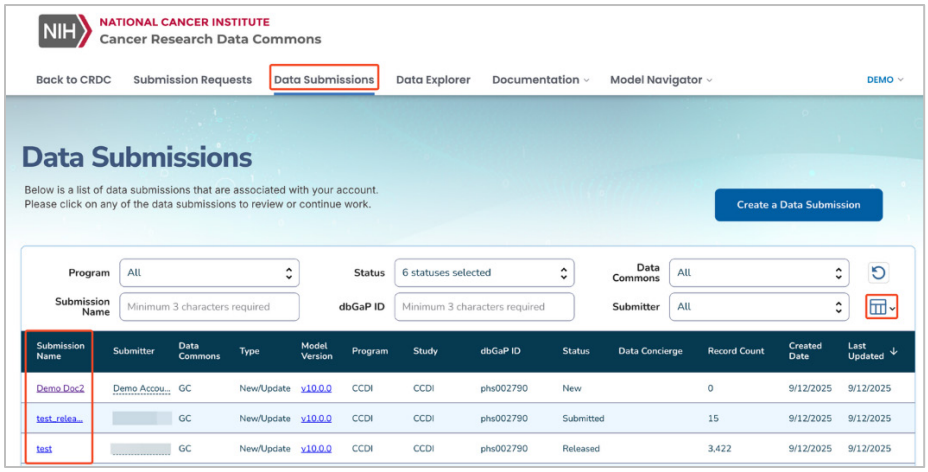

Figure 8. Active Data Submissions List

1. Adding and Managing Collaborators from the Dashboard

At the top of the new data submission page, the dashboard provides the tools to manage collaborators and the submission details. The submitter can add collaborators to a given submission and manage their access so that the collaborators can also upload and validate data through the CRDC Submission Portal.

- By default, the collaborators count is set to zero. As the collaborators are added, the count automatically updates on the dashboard. See Figure 9.
- The dashboard also displays the Submission ID, which is unique to each data submission.
- Icons next to the Study and Program allow submitters to copy the full name with a single click.
- The name and email of the data concierge assigned to the submission are also displayed on the dashboard.

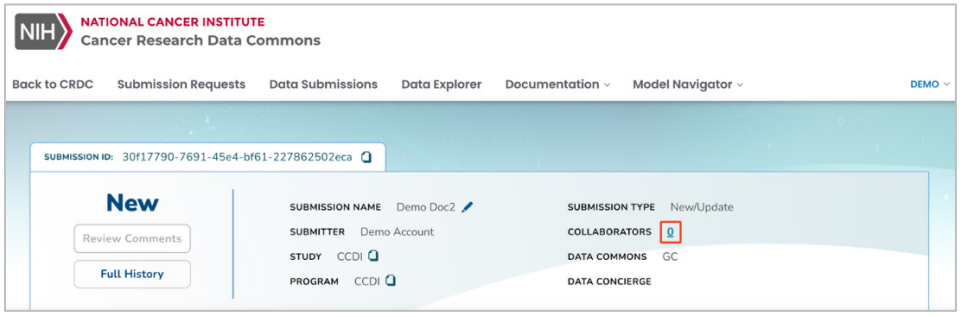

Figure 9. Count of Collaborators on the Data Submission Dashboard

To add and manage collaborators, click the hyperlink next to **Collaborators** on the dashboard, which opens a new Data Submission Collaborators window. See Figure 10.

Ensure that the collaborator has an authorized account on the CRDC Submission Portal with the Submitter role, is affiliated with the same study, and has permission from the data owner or Study PI(s) to submit data. (see "Request Access" on page 4). In the Data Submission Collaborators window, the submitter can select collaborator/s by selecting names from the dropdown list.

The submitter can **remove** a collaborator by clicking the remove icon [**X**] and add multiple collaborators by clicking the **Add Collaborator** button and repeating the process for each additional collaborator. The collaborators can upload and remove the data from the submission. Be sure to click **Save** before closing the window to retain any changes.

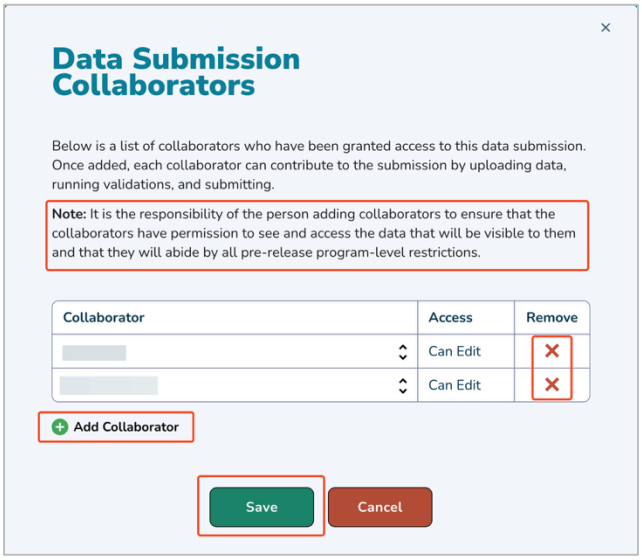

Figure 10. Add and Remove Collaborators for a Given Submission

2. Obtaining Submission Templates

Submitting data to CRDC requires you to submit associated metadata using the submission templates. The metadata is then used to validate the submitted actual raw data. For instance, the file name and the file size of each uploaded raw data file will be compared with the file name and file size specified in the metadata manifest.

To get to the submission templates, start with clicking the **Model Navigator** in the menu bar (see Figure 11), which lists the data models of the various Data Commons integrated under the CRDC Data Submission Portal. The data models of the various Data Commons, which include the General Commons (GC) Model, Clinical and Translational Data Commons (CTDC) Model, Integrated Canine Data Commons (ICDC) Model, and Population Science Data Commons (PSDC) Model, are listed. Models of other Data Commons will be added as and when the DCs are integrated with CRDC Submission Portal. Select the data model respective to the Data Commons (DC) to which the Submission Review Committee (SRC) has approved you to submit data.

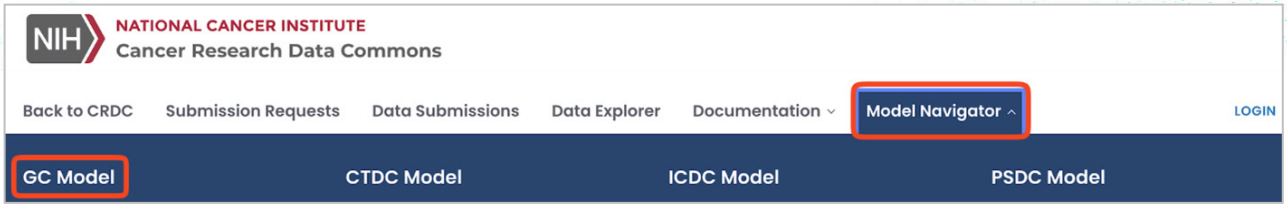

Figure 11. Use the Menu Bar to Navigate to the Data Model Viewer

Once you select the data model, you are taken to the Data Model Viewer page, as seen in Figure 12. On this page, you can view the data model in detail and download the submission templates from the dropdown list of Available Downloads. **Note:** Figure 12 shows the GC Data Model as an example.

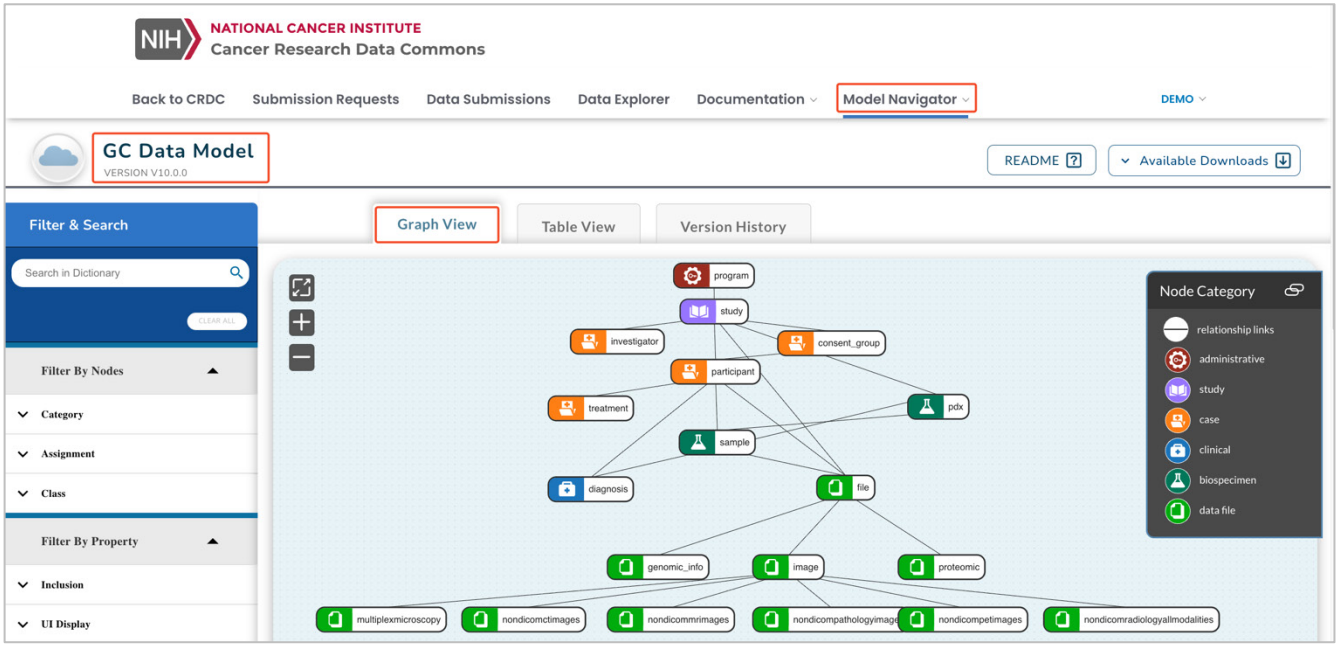

Figure 12. Data Model Viewer Graph View

Use the Data Model Viewer to explore the data elements that the relevant Data Commons requires or can accept. On the Graph View tab, the data model is represented in graphical nodes and relationships. Clicking a node in the graph shows its summary. At the bottom of the node summary, click **View Properties** to open a Table View of the selected node. For instance, as shown in Figure 13, clicking the **Diagnosis** node opens its summary, with the View Properties option at the bottom.

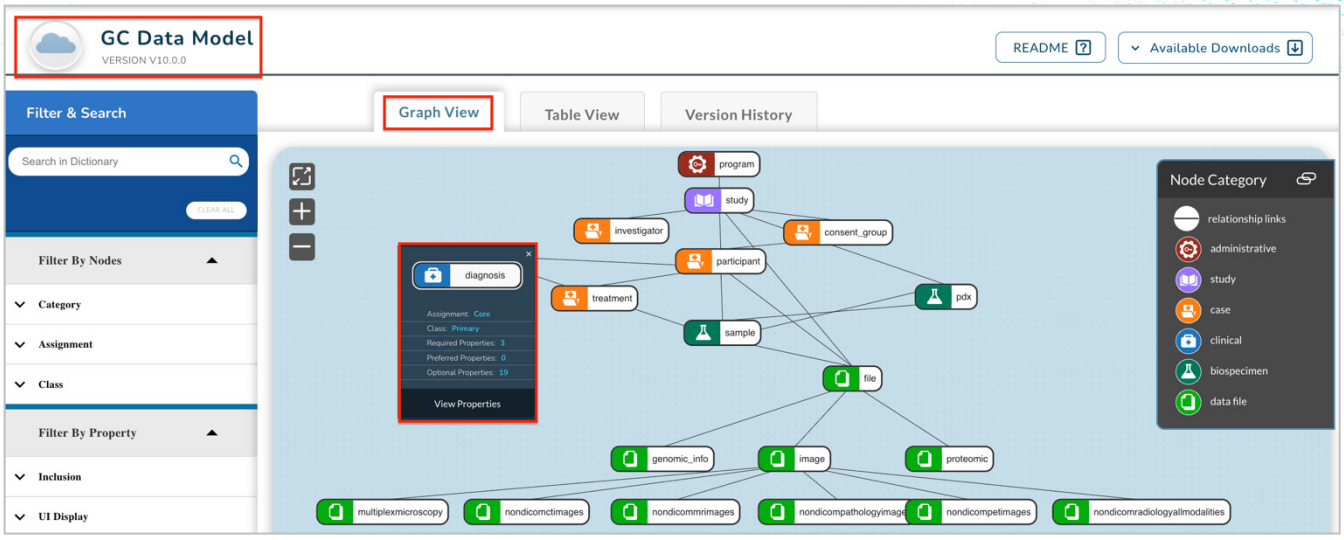

Figure 13. Click a Node to View a Summary and Open the Table View

The Table View lists all the data elements/properties of the data model and includes the description of each of these properties (such as strings, integers, etc.). The Table View also shows which of these properties are required for a submission. See Figure 14. Each of these properties is mapped to the Common Data Elements (CDEs) of the caDSR standards where applicable. Please note that CRDC data validations will only accept **Permissible Values** (PVs) for elements mapped to the **Common Data Elements (CDEs)**. Details about the CDEs can be accessed by clicking on the Public ID for that specific property. If you do not find the correct match in the provided values, you can request new CDEs and/or new PVs by emailing the CRDC Help Desk (NCICRDC@mail.nih.gov). These requests will go through a CRDC approval process before you can use the new CDEs or PVs in a submission.

Additionally, in the Table View, the Submission Template and the associated Data Dictionary for the specific node can be downloaded as shown in Figure 14 and described in the next section.

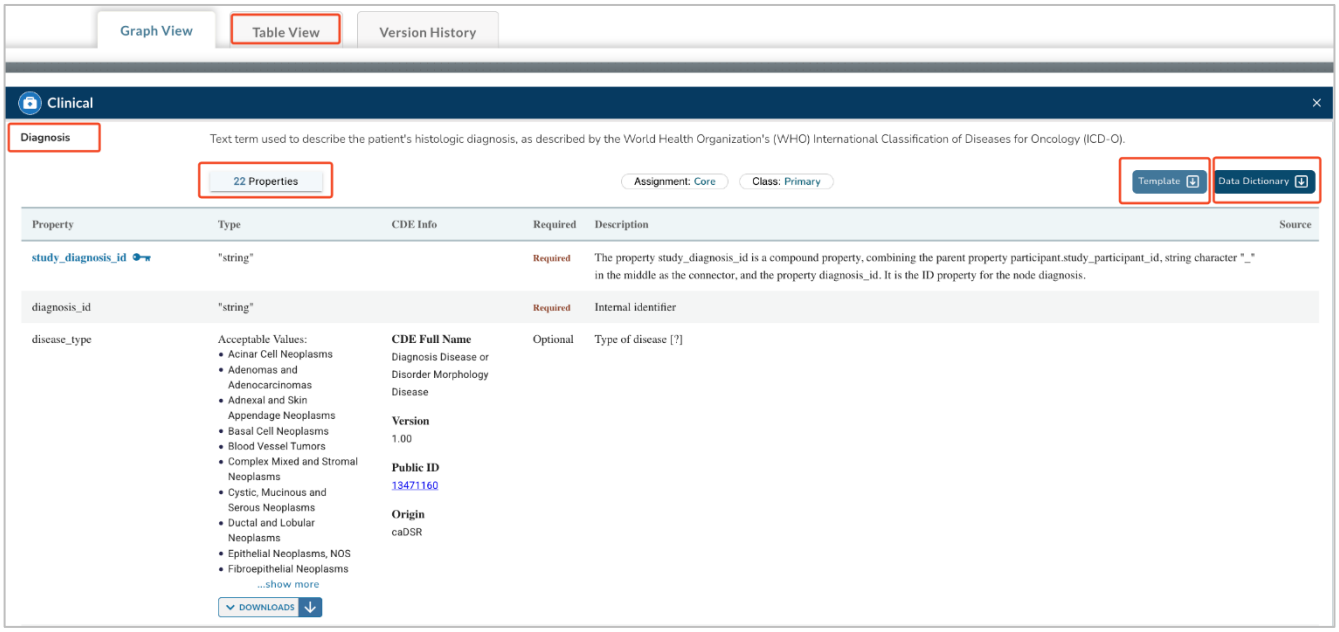

Figure 14. Table View of an Excerpt of Diagnosis Node

Additionally, any updates to the data model, such as property modifications, additions, or updates to permissible values, are reflected in the **Version History** tab, ensuring users have access to the changes. See Figure 15. Currently, Version History is applicable and available for the GC data model only.

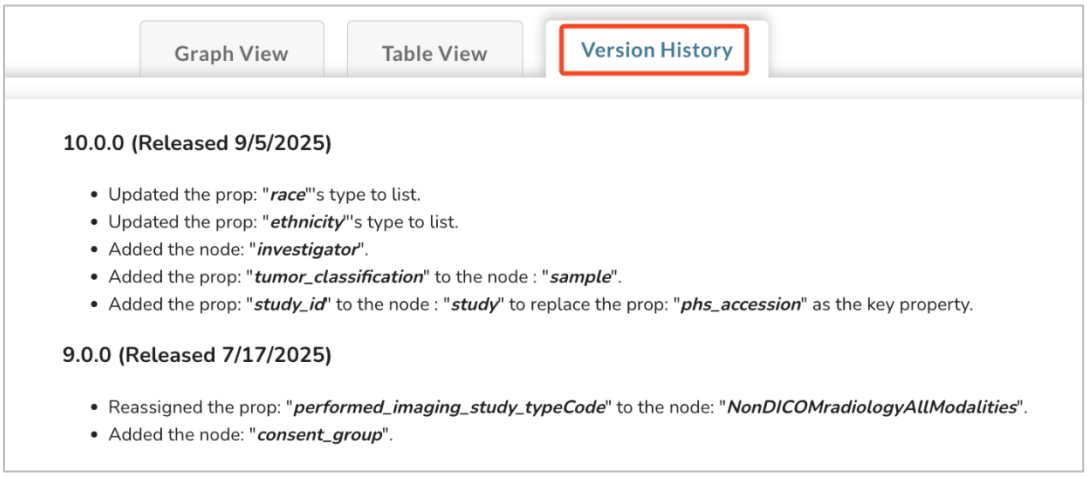

Figure 15. Excerpt of Version History of the GC Data Model

3. Downloading Data Dictionary and Submission Templates

Each Data Commons has its own unique data model with corresponding data dictionary, set of nodes, properties, and submission templates.

To download the Data Dictionary and the Submission Templates, select **Available Downloads**, as shown in Figure 16. Choose the options from the dropdown list to download the file in the preferred format, or download selected submission templates or all templates, as shown in Figure 17. You can also download CRDC Vocabularies for the selected Data Commons Model and Example Templates.

- Data Dictionary
  - All Properties (PDF, TSV, JSON)
  - Required Properties (PDF, TSV, JSON)
- Submission Templates (TSV)
  - All Templates
  - Selective Templates
- All Vocabularies (TSV, JSON)
- Example Templates

Submitters can use the **Submission Templates** to format and upload metadata to the CRDC Submission Portal. These templates ***must*** be in TSV format. Templates in any other format, such as Microsoft Excel (.xls, .xlsx) etc., will fail. You can use software such as ModernCSV to work with these submission templates, as it handles CSV and TSV as tables without automatically modifying the data.

The **Data Dictionary** provides detailed information about the metadata structure, content, and the required, preferred, and optional data elements for all nodes within the selected data model. Submitters can choose to download the Data Dictionary for either all properties or only the required properties. The

All Vocabularies document contains the permissible values for the data elements. The **Example Templates** are examples of completed submission templates with mock data, designed to guide users in preparing the metadata manifest for their data. These can be useful to understand what each of the columns in the template is supposed to contain.

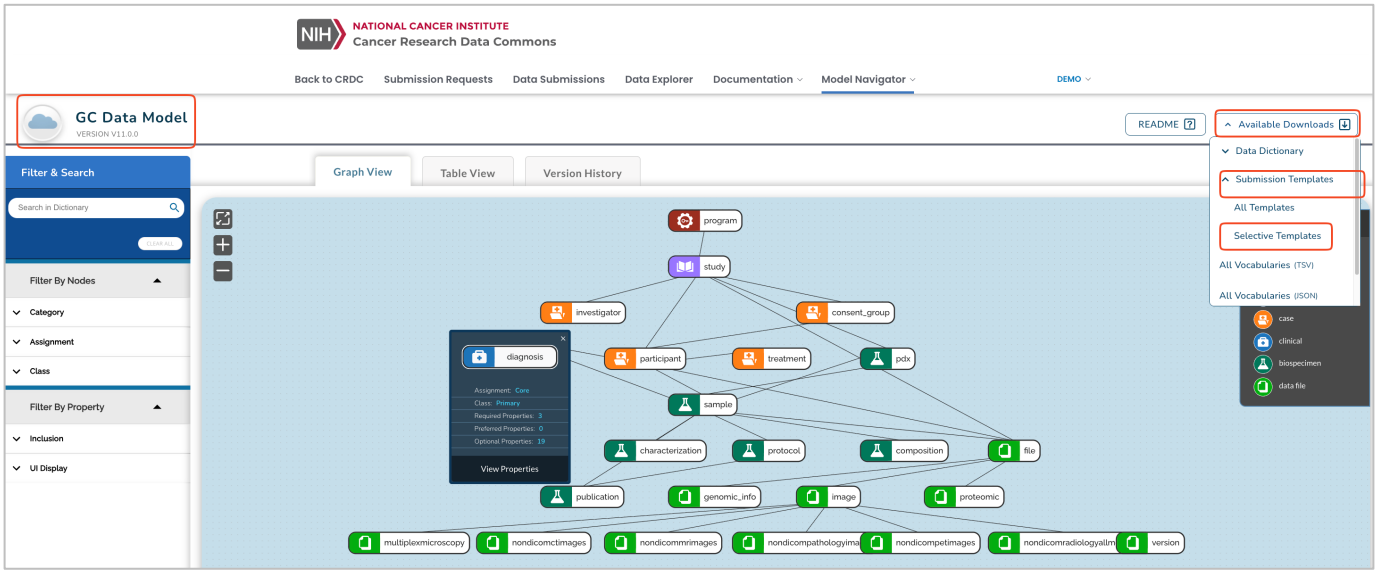

Figure 16. Using the Available Downloads Menu

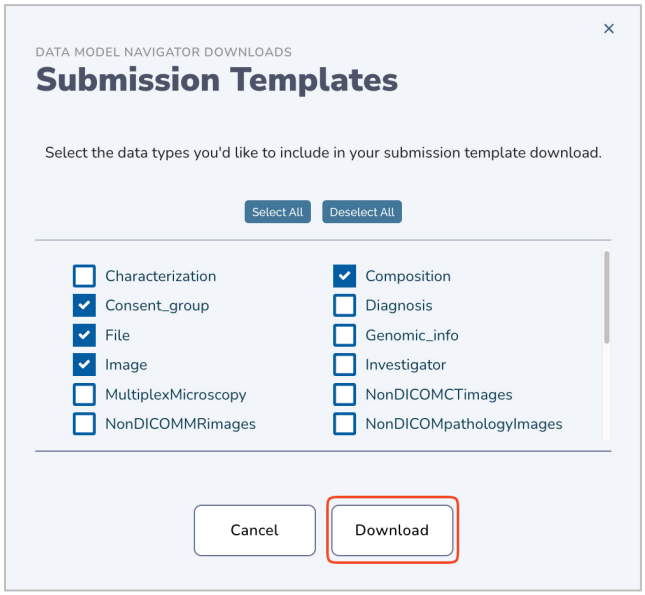

Figure 17. Download Selected Submission Templates

The downloaded files are provided as a ZIP archive. The tab-separated text files can be viewed in any text editor or spreadsheet application like Microsoft Excel or OpenOffice Calc. Multiple metadata template files in TSV format are included within the ZIP archive called Submission Templates. The downloaded and unzipped All Submission Templates folder for the GC model is shown in Figure 18.

Note: The exact content of the submission templates differs depending on the selected data model and the associated submission process requirements.

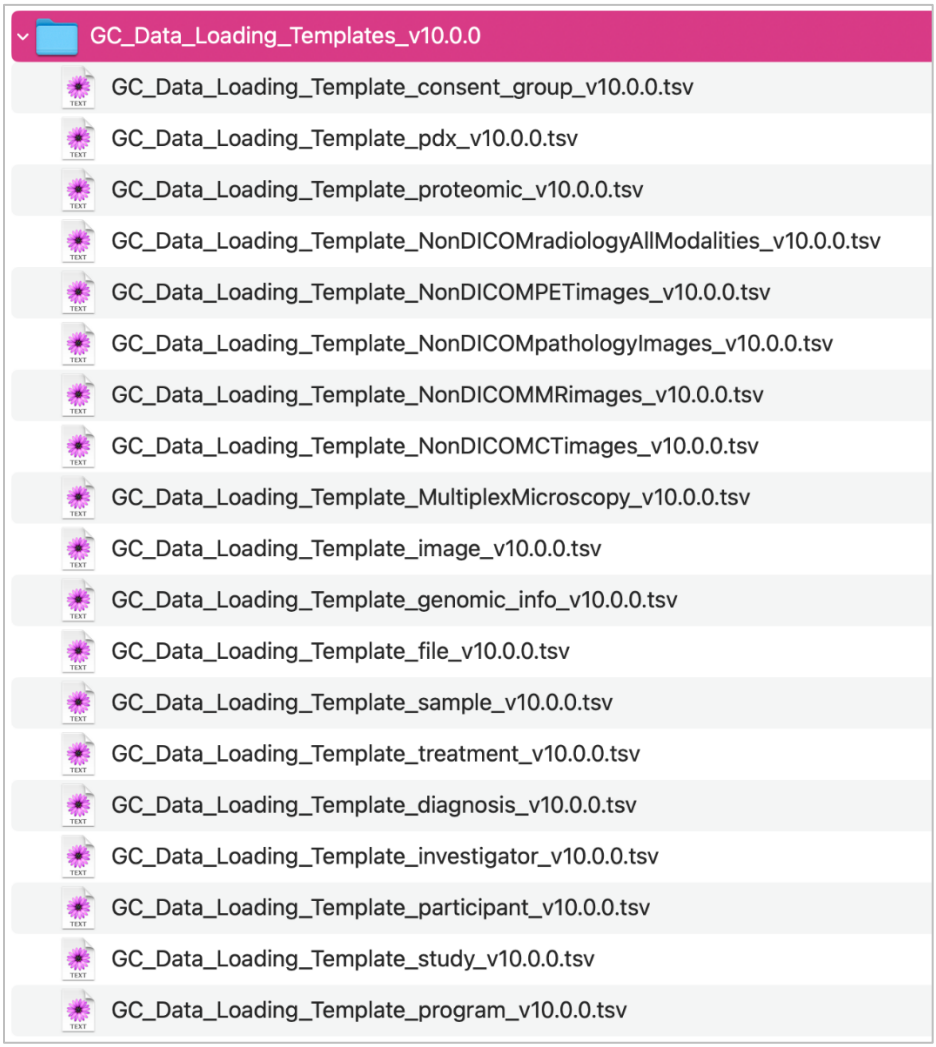

Figure 18. Submission Templates Downloaded from the CRDC Submission Portal Model Viewer

3.1 Submission Templates and Properties

Each of the submission templates covers information relevant to a specific node in the model; for example, the template, 'GC_Data_Loading_Template_image_v10.0.0.tsv,' collects imaging data related information. Only those templates relevant to the data being submitted are required. For instance, if the submission does not include imaging data, the submitter does not need to fill or submit the 'GC_Data_Loading_Template_image_v10.0.0.tsv.'

For every template that will be submitted, review the Data Dictionary (accessible through the **Available Downloads** menu) to understand which properties are required, as each individual template has required properties, as well as preferred and optional properties in the node. Note that you should not edit the first row of each template as it contains the property names and other special columns, explained below.

3.1.1 Special Columns

Every template has two types of special columns, also called parent mapping columns: “type” and “relationship.”

3.1.2 Type Column

A “type” column contains the name of the node type (such as study or genomic_info) and is required by all template files. In the downloaded template, the second row in the first column is pre-populated with the correct node name for that specific template (e.g., in the ‘GC_Data_Loading_Template_study_v10.0.0.tsv’ template, the second row in the first column is filled in as ‘Study).’. All rows should contain the same node name in the “type” column. Mixing multiple node types in one file is not supported. For example, the node type ‘Sample’ should not be mixed with the node type ‘File.’

3.1.3 Relationship Columns (Parent Mapping Columns)

A “relationship” column is used to specify relationships between the current node and its related nodes. A relationship column has a header in the form of “<parent node name>.<parent ID property name>.” Values in the relationship columns are IDs of the related nodes (like a foreign key in a relational model).

For example, for the study node, the “program.program_acronym” column indicates that the study node has “program” node as its parent node, and the property used to identify the program node is “program_acronym.” Each value in the “program.program_acronym” column is an acronym used for a program, such as HTAN.

4. Uploading Data Files and Metadata Manifests

You can move files from your local environment to the CRDC through the Submission Portal in the following two ways:

- **Uploader CLI Tool** – This command-line interface is used to transfer primary data files like genomic sequence files or imaging data files to CRDC.
- **Graphical interface** – The graphical interface can be used to upload metadata files such as the Submission Templates.

Note: Submit primary data files using the Uploader CLI Tool only. Do not attempt to upload data files using the CRDC Submission Portal’s graphical interface. Submitters can choose to upload the supplementary files, if requested by the CRDC Submission Team, using the CLI tool. However, the validations will not be performed on the supplementary files.

4.1 Uploader CLI Tool

4.1.1 Introduction

The CRDC Submission Portal provides a command-line interface (CLI), called the Uploader CLI Tool, for uploading data to its temporary CRDC storage. You can install and use the Uploader CLI Tool on any system capable of running Python 3.6 or higher. Binary versions of the Uploader CLI Tool are also available, which don’t require any installation or Python.

Notes:

- There are detailed instructions on downloading, installing, and running the Uploader CLI Tool in the README file of the GitHub repository.
- The Uploader CLI Tool does not have to be downloaded for each submission; this is a Python script that can be used for any upload to the CRDC Submission Portal. The only aspect that must be tailored to each submission is the configuration file, which is discussed below. However, submitters should ensure that they are using the **latest version** of the Uploader CLI Tool and the configuration file.

4.2 Downloading the Uploader CLI Tool

You can download the Uploader CLI Tool either directly from the CRDC Submission Portal or by cloning the GitHub repository. Downloading from the CRDC Submission Portal is recommended as it ensures you are using the latest version. Starting from version 4.0, the Uploader CLI Tool will automatically check if your version of the Uploader CLI tool is compatible with CRDC Submission Portal backend, and prompt you to download a new version when it is no longer compatible.

4.2.1 Download the Uploader CLI Tool from the CRDC Submission Portal

Open the menu by clicking your user profile name, found in the upper-right corner of the Data Submission page. See Figure 19. Select **Uploader CLI Tool** from the menu to open a pop-up window with the latest version of the CLI tool.

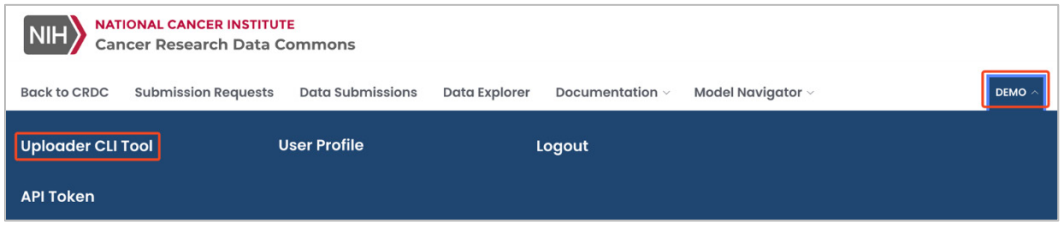

Figure 19. Menu with the Uploader CLI Tool Download Option

Click the **Download** icon next to the available Package Type options to download the Uploader CLI Tool. The download comes with accompanying instructions (see Figure 20). A ZIP archive will be saved to your local machine.

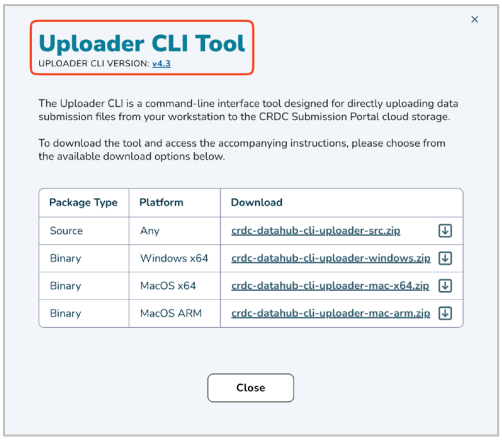

Figure 20. Download the Uploader CLI Tool

4.2.2 Cloning the Uploader CLI Tool from GitHub

The latest version of the Uploader CLI Tool can also be cloned from the Data Hub GitHub repository (see Figure 21). To clone the repository to your local machine, use the following command:

```
git clone --recurse-submodules
https://github.com/CBIIT/crdc-datahub-cli-uploader.git
```

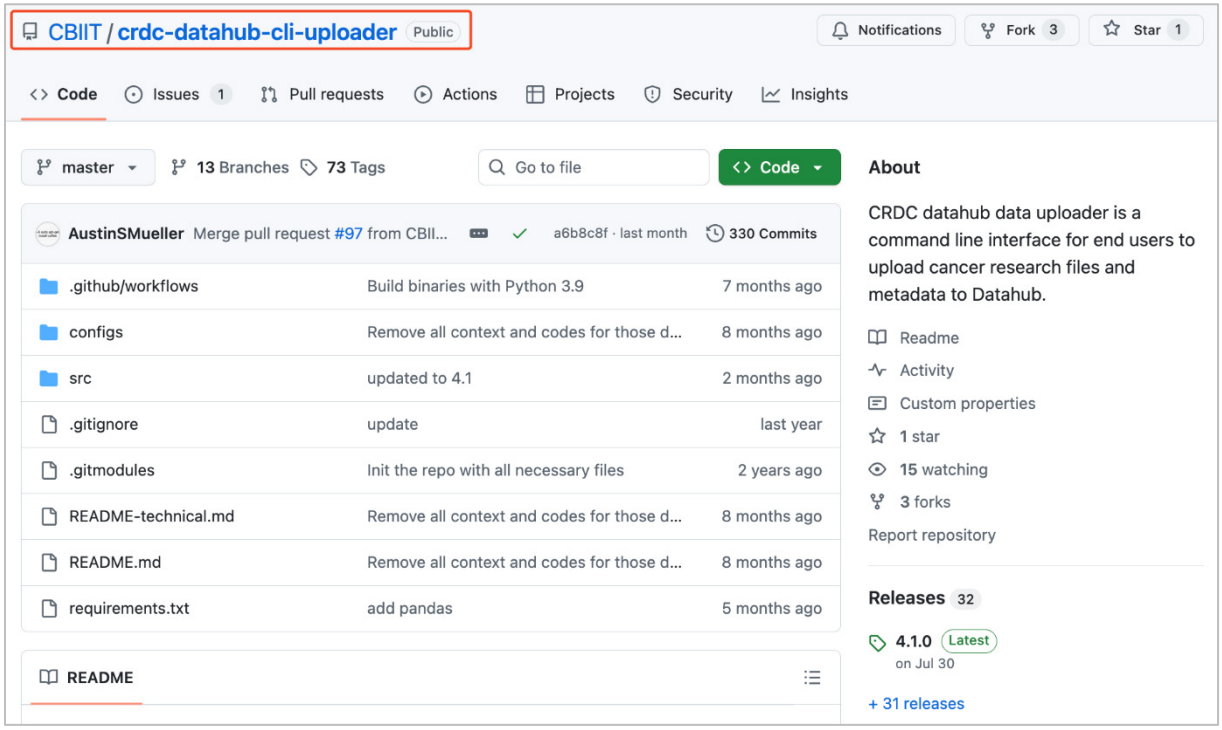

Figure 21. Uploader CLI Tool as it Appears in GitHub

4.3 Setting Up the Python Environment for CLI (Source version)

Binary versions of the Uploader CLI Tool are self-contained and don't require Python or installation of any dependencies. The Source version of Uploader CLI Tool has Python library dependencies that you must install before running the CLI. These dependencies can be installed by running the command `pip3 install -r requirements.txt`. The *requirements.txt* file contains the list of dependencies, described below. If you want to install the dependencies individually, install the following libraries:

- pyyaml
- boto3
- requests
- requests_aws4auth
- Rich
- pandas

4.4 Using the Uploader CLI Tool

Using the Uploader CLI Tool to upload data to Submission Portal requires preparing a few types of files that are described in the following subsections.

- **Configuration File** (YAML format). See detail in “Configuration File in YAML Format” on page 18.
- **File Manifest** (TSV format). See detail in “File Manifest in TSV Format” on page 21.
- **Archive Manifest, if applicable** (TSV format). See detail in “Archive Manifest in TSV Format” on page 22.

4.4.1 Configuration File in YAML Format

You can locate the configuration file through the Submission Portal or from the CLI folder.

Finding the Download Configuration File button on the Submission Portal: The configuration file controls the behavior of the **Uploader CLI Tool**. You can directly download the configuration file from the CRDC Submission Portal by clicking the **Download Configuration File** button, which is shown in Figure 22. You can also access the Data Submission Instructions document, the version of the data model your submission uses, and the Uploader CLI Tool on this same page.

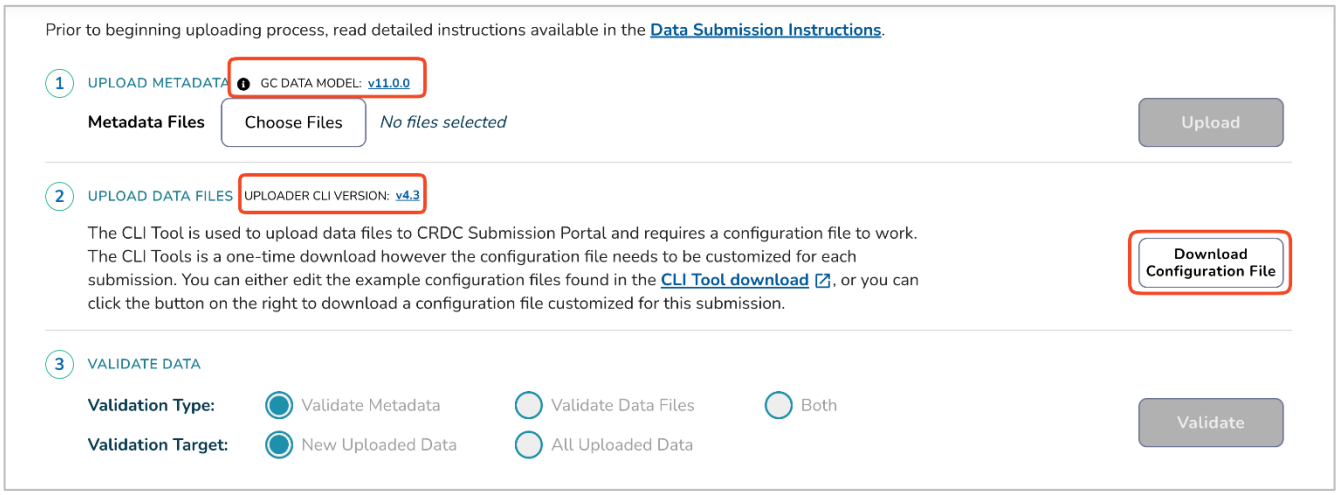

Figure 22. Download Configuration File

After you click the **Download Configuration File** button, the **Download Configuration File** pop-up window appears, as shown in Figure 23.

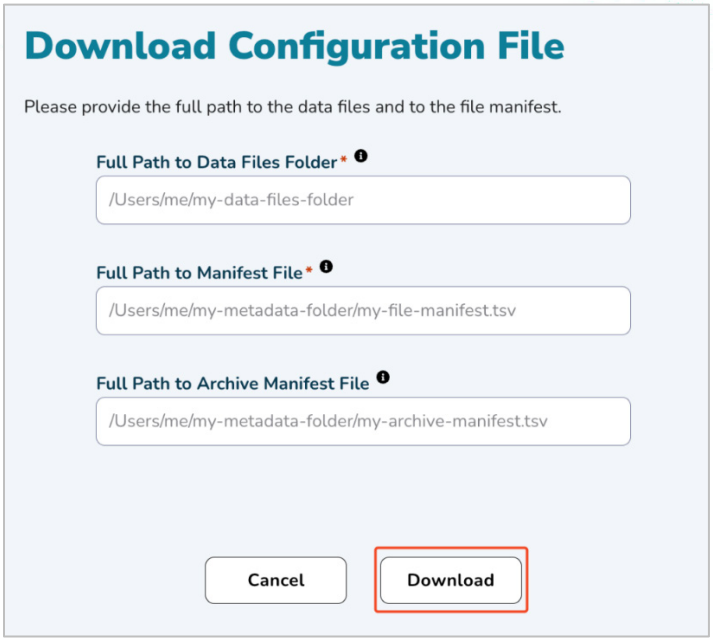

Figure 23. Dependencies to Download Configuration File

Dependencies for downloading the configuration file:

- **Full path to the data file folder:** Enter the path to the local or S3 folder containing your data files.

  Please note that if your data files are organized in a nested folder structure, the Uploader CLI Tool will automatically rename the data files to avoid filename collision. To correctly upload data files from nested folder, provide the relative path to each data file in the file_name property of the file manifest.

  For example, also shown in Figure 24, if data files are stored in a folder named "data files" and the file is located at data files/folder1/folder2/abc.bam, then you should enter "folder1/folder2/abc.bam" in the file_name column – not just abc.bam.

  As a result, the uploaded file will be renamed to folder1_folder2_abc.bam in the CRDC storage. However, the file_name property in the **final_manifest** will still reflect the original file name, abc.bam.

  Note: If you enter an S3 URL in the **full path to Data Files folde**r, the Uploader CLI Tool initiates an S3-to-S3 transfer.

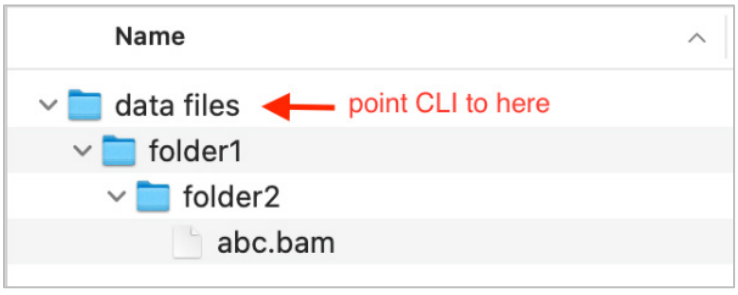

Figure 24. CLI Tool Requires Relative Path to Data File in a Nested Directory

- **Full path to the File Manifest**. Enter the full path in the designated text box. File Manifest is explained in “File Manifest in TSV Format” on page 21.
- **Full path to the Archive Manifest, if applicable:** If you prefer to upload the files as a ZIP archive, you must provide the path to the **archive file manifest.** This is explained in “Archive Manifest in TSV Format” on page 22.
- **Downloading the Configuration file:** Clicking the **Download** button will download the configuration file in YML format to your computer with pre-populated fields.

Editing Configuration file manually: Additionally, if you choose to populate the configuration file manually, you can find examples in the *config* directory of either the extracted ZIP file or the cloned GitHub repository. The examples provided are the same configuration file modified for the two different upload types.

- **Uploader-metadata-config.example.yml** – This file is an example of the configuration file needed by the Uploader CLI Tool to upload metadata submission templates rather than submitting them via the CRDC Submission Portal graphical interface.
- **Uploader-file-config.example.yml** – This is an example of the configuration file needed by the Uploader CLI Tool for uploading large primary data files such as BAM files. Files uploaded this way go through the file validation system rather than the metadata validation system.

Description of the fields in the configuration file: These configuration files are in YAML format and the Uploader CLI Tool will fail if the file is not a valid YAML. YAML-aware text editors such as Microsoft Visual Studio Code, Sublime Text, or Notepad++ can be extremely helpful in preserving YAML formatting. The fields in this file follow.

- **api-url** – This field provides the Uploader CLI Tool with the URL/location of the temporary CRDC storage used for API communications and upload.
- **token** – This is the API access token that is obtained from the CRDC Submission Portal’s graphical interface. To obtain an API token, log into the CRDC Submission Portal graphical interface to bring up the user menu, then select **API Token**. This opens a dialog box that allows you to create and copy an API token to your clipboard.
- **submission** – This is the submission ID that identifies which study the uploaded files will be associated with. To find the correct submission ID, log into the system and select the study from the Data Submissions List by clicking the submission name. You can copy the Submission ID from the upper-left corner of the interface by clicking the icon to the right of the Submission ID number.

  Note: A study consists of one or more submissions (often many more), with each Submission ID linked to the parent study. A single user working on multiple studies must carefully track which Submission IDs they are uploading to ensure the data is associated with the correct study.
- **type** – This tells the system if this is a metadata upload or a data file upload. Enter the term *metadata* if the upload contains submission templates and *file* if the upload contains data files.
- **data** – This is the local path to the directory that contains the files to be uploaded.
- **manifest (*Data file upload only*)** – This is the local path to the file manifest.

- **archive_manifest** (required for ZIP archives) – This is the path to the archive manifest. This file is required for uploading any zipped (.zip) data files
- **retries** – This is the number of retries the Uploader CLI Tool will perform after a failed upload.
- **overwrite** – If this is set to *true*, the Uploader CLI Tool overwrites the file with the same name that already exists in the CRDC Submission Portal target storage. If set to *false*, the Uploader CLI tool does not upload if a file with the same name and size exists in the CRDC Submission Portal target storage.
- **dryrun** – If this is set to *true*, the Uploader CLI Tool does not upload any files to the CRDC Submission Portal target storage. If set to *false*, CLI uploads files to the CRDC Submission Portal target storage.

While users are expected to provide paths to their data folder and file manifest, they may choose to customize the values of the three parameters—**retries**, **overwrite**, and **dryrun**—to suit their needs.

4.4.2. File Manifest in TSV Format

The Uploader CLI Tool uses a document called a file manifest to upload the data files to the temporary CRDC storage. The file manifest is a simple table (a TSV file) with all the required properties as defined by the data model except the file IDs, which are generated by the Uploader CLI Tool. Submitters can use the file.tsv template downloaded from the Data Model viewer page, to create this file manifest, saving the effort of creating a duplicate file.

Instructions for using the file.tsv template with the Uploader CLI Tool to upload data files:

1. Download the file.tsv template from the Data Model Viewer page. The file.tsv file does not include a column for file IDs/Keys because the IDs are generated by the Uploader CLI tool and automatically embedded into the file.tsv during the data upload process.
2. Also download templates for child nodes which have empty file ID columns.
3. Populate the templates to prepare the file.tsv manifest (template populated with metadata) and the manifests for child node that include file ID columns
    - Ensure these child node manifests **do** include columns for file IDs.
    - Ensure correct file names are filled in columns for file IDs. The Uploader CLI Tool will generate final manifests for child nodes and replace the file names with corresponding file IDs.
4. Organize files.
    - Place the file.tsv and all child node metadata manifests into the same folder.
5. Run the Uploader CLI Tool.
    - Use the CLI tool to upload the data files.
    - During the upload the tool processes the manifests and generates file IDs.
6. Generate the final version of manifests.
    - Once the data upload is complete, the CLI tool automatically creates the “final” versions of the manifests (e.g., file-final.tsv).
    - These final manifests are saved in the same folder as the original manifests.

7. The final manifests are uploaded to the Submission Portal by the Uploader CLI Tool, so users do not need to make changes in the final manifests.
    - If you need to make changes to the final manifests after the Uploader CLI Tool has uploaded them to the portal, ensure you edit the final version (e.g., file-final.tsv saved locally on your computer) and then re-upload it through the user interface and not the Uploader CLI Tool.
    - If you need to re-upload your data files for any reason, you can reuse the final manifest generated by the Uploader CLI Tool. This will retain the existing file IDs/keys instead of generating new ones.

4.4.3. Archive Manifest in TSV Format

This file is required when submitter has the data structured in ZIP archives and would prefer to upload them in the same format. The archive manifest lists the contents of the ZIP archive so that CLI can verify the contents before uploading them to Submission Portal. An example archive file manifest is shown in Figure 25. The Submitter can find the archive manifest template in the **config** folder of CLI. The archive manifest must include the following four columns:

- Archive_name: the name of the ZIP archive
- File_path: Full path to each file within ZIP archive
- File_size: size of each file in bytes
- Md5: MD5 hash of each file within the ZIP archive

	archive_name	file_path	file_size	md5
2	zipped-data.zip	level1/TCL01_CellLine_CDS_Variants.csv	4173475	B6CB025A2062A8F88F17E6314E6BE44C
3	zipped-data.zip	level1/level2a/TCL01_Study_Protocol.docx	219492	148ab1ba4faf414842cbf64750cd3a7a
4	zipped-data-2.zip	level1/TCL01_CellLine_CDS_Variants.csv	4173475	b6cb025a2062a8f88f17e6314e6be44c
5	zipped-data-2.zip	level1/level2a/TCL01_Study_Protocol.docx	219492	148ab1ba4faf414842cbf64750cd3a7a

Figure 25. A Sample Archive Manifest

4.5 Starting the Upload Process

Once the configuration file has been downloaded or edited, you can start the upload script. The only required parameter is `--config`, which should provide the full path and file name for the completed configuration file. The command should look something like the following, though the exact details may be customized depending on how the tool (and Python) were installed.

The following commands assume that your terminal's current working directory is set to the unzipped CLI folder.

When running the source code, use the following command:

```
$ python3 scr/uploader.py --config path/to/cli-config-my-submission.yml
```

When running the Windows version, the command should look like the following (note: do not try to open or double click uploader.exe file):

```
$ uploader.exe --config path/to/metadata-upload.yml
```

When running Mac version, the command should be (note: do not try to open or double click uploader file):

```
$ ./uploader --config path/to/metadata-upload.yml
```

4.6 Using the CRDC Submission Portal's Graphical Interface to Upload Metadata Submission Templates

The Upload Metadata feature in the CRDC Submission Portal's graphical interface is intended for submitting completed metadata templates. Users can access tooltips and a link to the data submission user guide directly from the interface. The Upload Activities table tracks the uploading process of data and metadata files and provides details on any errors related to failed uploads. After the data upload is complete, the system automatically runs the basic validation process, and results are shown in the Validation Results table. You can delete the specific data or metadata files for a submission from the Data View table.

Batch Uploads: If you intend to upload the metadata in batches, you should keep your associated metadata separated by participants. For example, if a study has 100 participants, the submitted template for the first batch could either contain all 100 or a subset of that 100, with the remainder submitted in later batch uploads. If there is no overlap in participants between the different uploads, the system will not flag an error. However, mixing new data from previously uploaded participants with new participants will result in an error, as the system knows about the previously uploaded participants. To make corrections, select the file you want to delete in the Data View table and click the **Delete** button. Note: If the system detects an issue even in one file within the batch, the entire batch will fail, and the Submitter will need to address the errors before reuploading.

Process of uploading Metadata via GUI: As depicted in Figure 26, to start the upload process, click the **Choose Files** button and then select the metadata submission manifests you want to submit. The total number of files that you have selected appears. If that number is correct, click the **Upload** button to start the upload. The Status column in the Upload Activities table displays *Uploading* until the upload and basic checks are completed. Successful files show as *Uploaded* in the Status column. If a file fails the basic checks performed by the Uploader CLI Tool, the data will not be uploaded and the status will display as *Failed*.

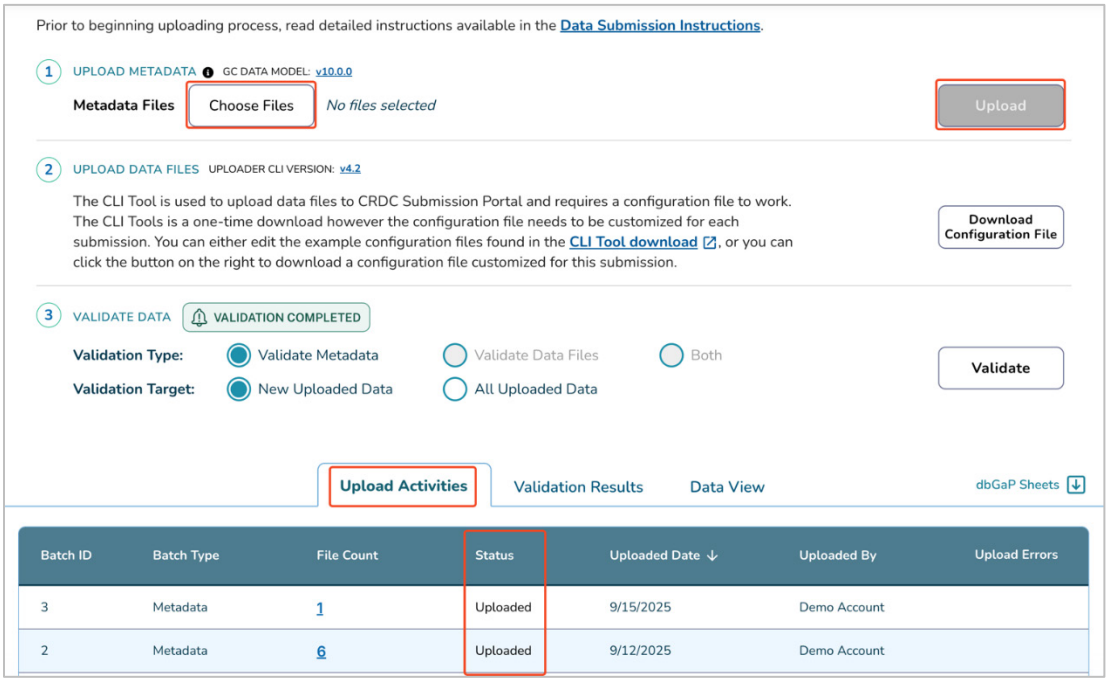

Figure 26. Upload and Validate Data and Metadata

Upload Activities Log: Clicking the number under the File Count column displays a list of all files uploaded in that batch. If there are errors in the files being uploaded, the Status column displays *Failed* and the Upload Errors column displays a link to the errors. Clicking that link opens a dialog box that explains what errors have been encountered. Correct all identified errors and reupload the file(s). If multiple files are uploaded in a batch, a failure in one of the files fails the entire batch. All files in a failed batch must be reuploaded. An example of a batch upload error message is presented in Figure 27.

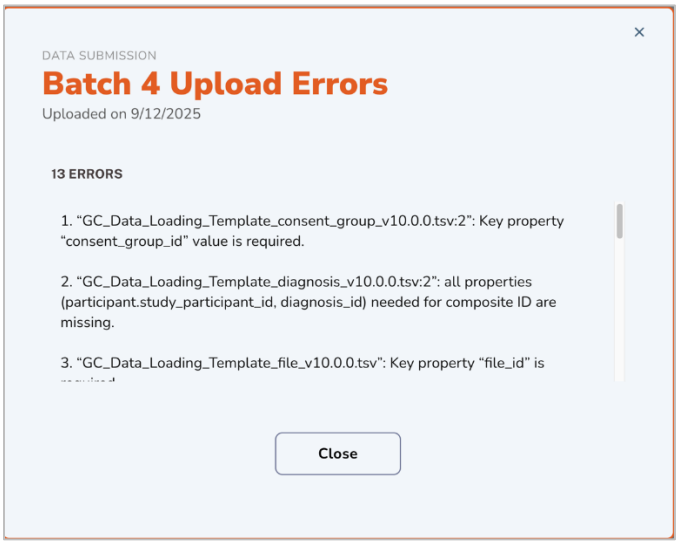

Figure 27. Batch Upload Errors

Data View Table Features: You can remove specific metadata files previously uploaded in the system in the Data View table, as depicted in Figure 28. For instance, to remove metadata for a participant, select the file and click the **Delete** icon to remove it from your submission. Note, this action will also delete the associated metadata from the child nodes. You can also download the selected metadata files from the Data View table by selecting the file(s) and clicking the cloud-shaped **download** icon. See Figure 29.

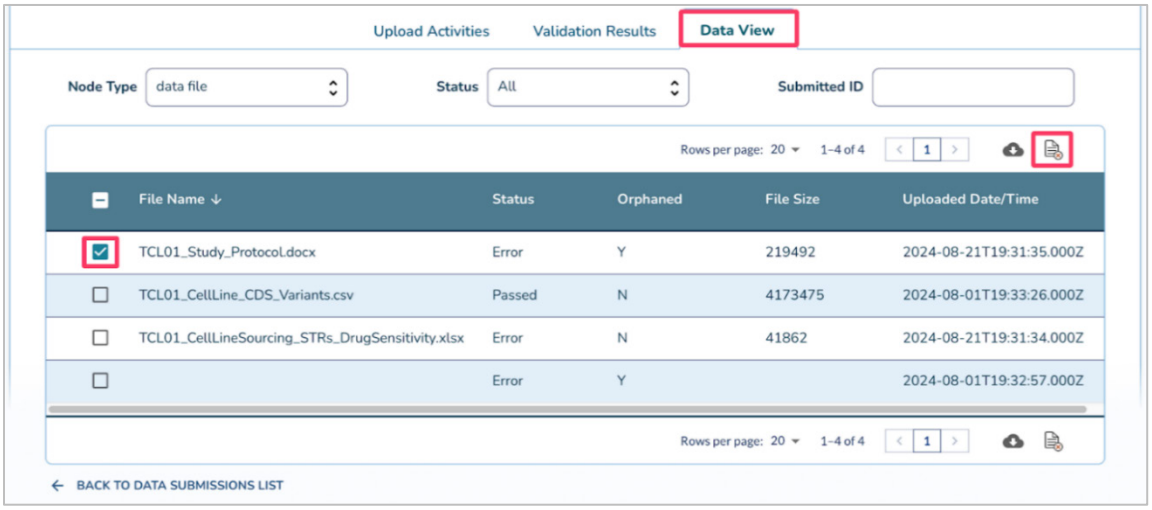

Figure 28. Delete and Download Metadata File(s)

View the associated metadata from the child nodes: Click the File Name in the first column of the Data View table. This action opens a window displaying metadata linked to the selected item. For example, clicking on the file name *TCLO1_Study_Protocol.docx* opens a window like the one in Figure 27. The window displays the related metadata from the child nodes such as Study, Participant, Diagnosis, and Specimen. Note that the specific child nodes may vary depending on the data model used by each CRDC Data Commons.

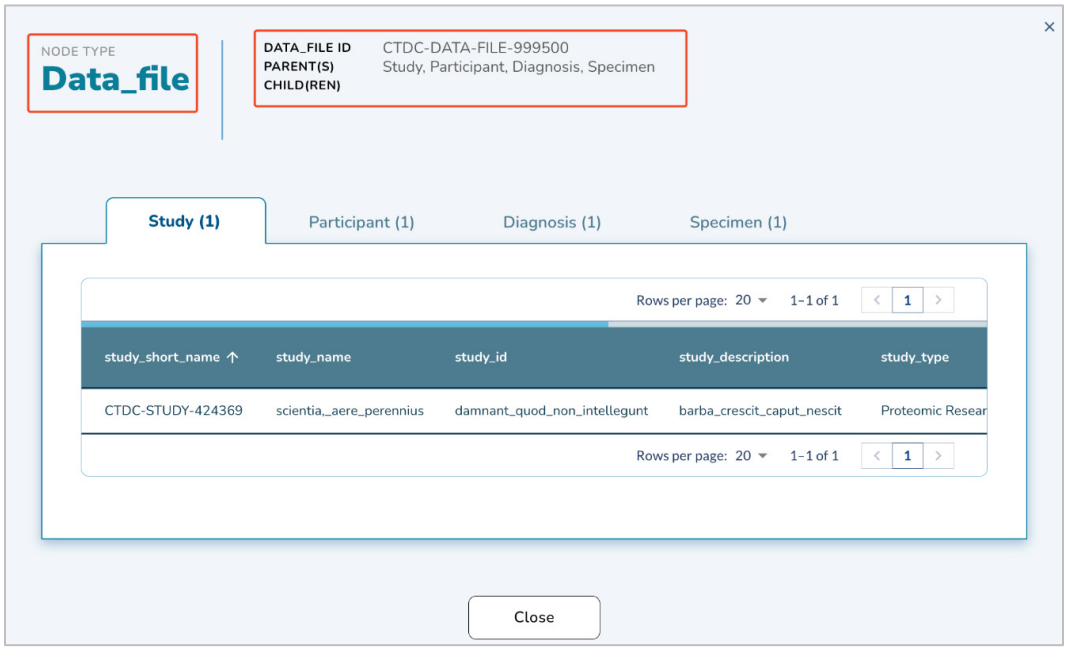

Figure 29. Visualize Associated Metadata from Child Nodes

5. Running Validations

Validations can be run at any point in the submission process, without restriction on when or how often. To run a validation, select options in the Validate Data panel and then click the **Validate** button. Refer to Figure 30.

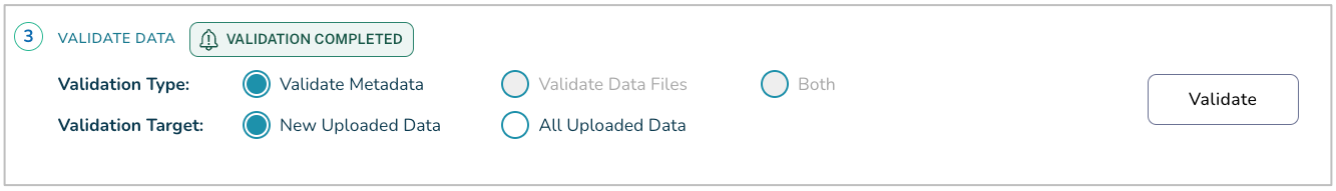

Figure 30. Validate Data Options

The first step is selecting which files to validate. The **Validate Metadata** option runs validations only on the submitted Metadata Manifest, not on any of the uploaded data files. The **Validate Data Files** option does the reverse and checks all the uploaded data files. The **Both** option validates both.

By default, only newly uploaded files are validated. This can be a significant time saver for large submissions as some validations take considerable time and the system keeps a record of previously submitted files that have already passed validation. However, if there is a need to check the entire submission, regardless of previous validations, the **All Uploaded Data** option validates everything that has been uploaded so far.

5.1 Reviewing Validation Results

After validations are run, the graphics on the page are updated to give a summary of the results as depicted in Figure 31.

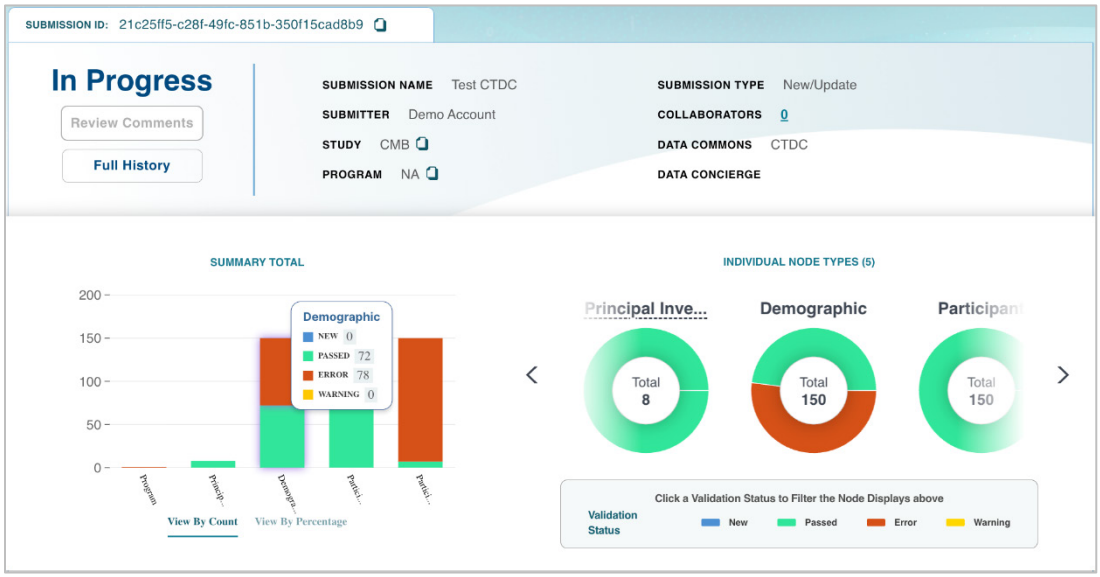

Figure 31. Validation Summary

The left graph in Figure 31 displays a list of the nodes and indicates the status of the uploaded and validated data, with green representing data that passed the validations and red representing data that failed the validations. Hovering over each bar generates a more detailed summary for that node. For instance, Figure 31 shows that the demographic node has 150 uploaded files, 72 of which successfully passed validation and 78 of which have errors. The blue color indicates data that have been uploaded but have not yet been validated. The graphs on the right are a node-by-node description of the results with the left and right arrows moving between the nodes that have been submitted to date.

4.1.1. Viewing and Filtering Validation Results

Submitters can view validation results in Aggregated (default) or Expanded format by clicking the toggle in the top left. Aggregated format groups validation results by Issue Type as shown in Figure 32.

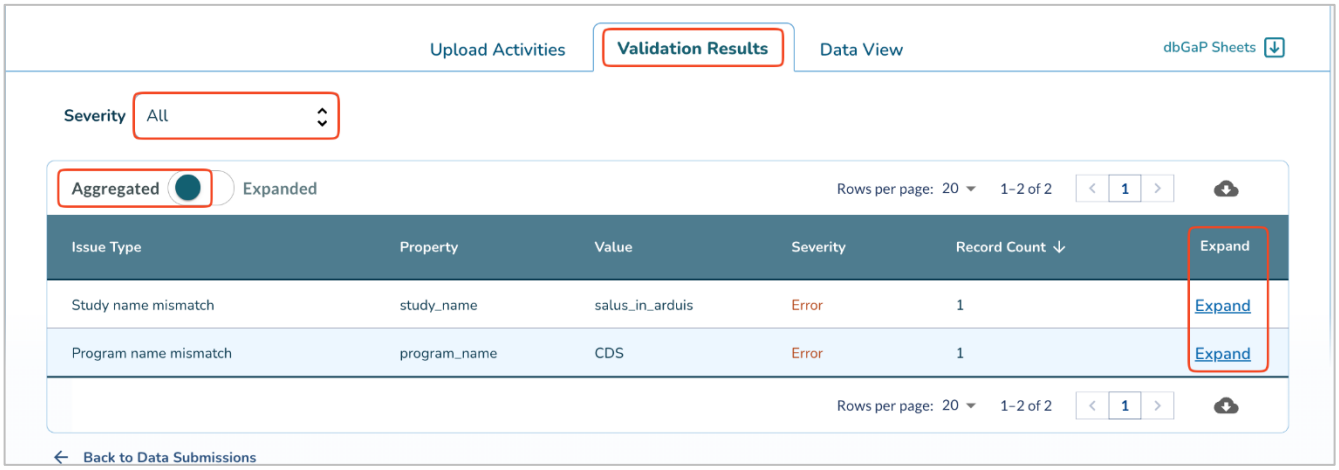

Figure 32. Aggregated Validation Results Table

To explore errors shown in the Expanded format (see Figure 33) in more detail, follow these steps:

a. Click the **Expand** link under the Expand column to view detailed validation errors by Issue Type.

b. Additional filters (Issue Type, Batch ID, Node Type, and Severity) can be helpful to narrow down the results. Selecting a specific **Node Type** refines the search to relevant validation errors. Choosing **All** from this menu displays a list of all files containing errors or warnings. The Node Type *file* pertains to metadata, while the *data file* refers to raw data, such as sequencing data.

c. View details for a specific Issue under the Issue column by clicking the **See details** link to access more specific information regarding each validation error.

d. Users can download all the validation errors in a table by clicking on the cloud icon.

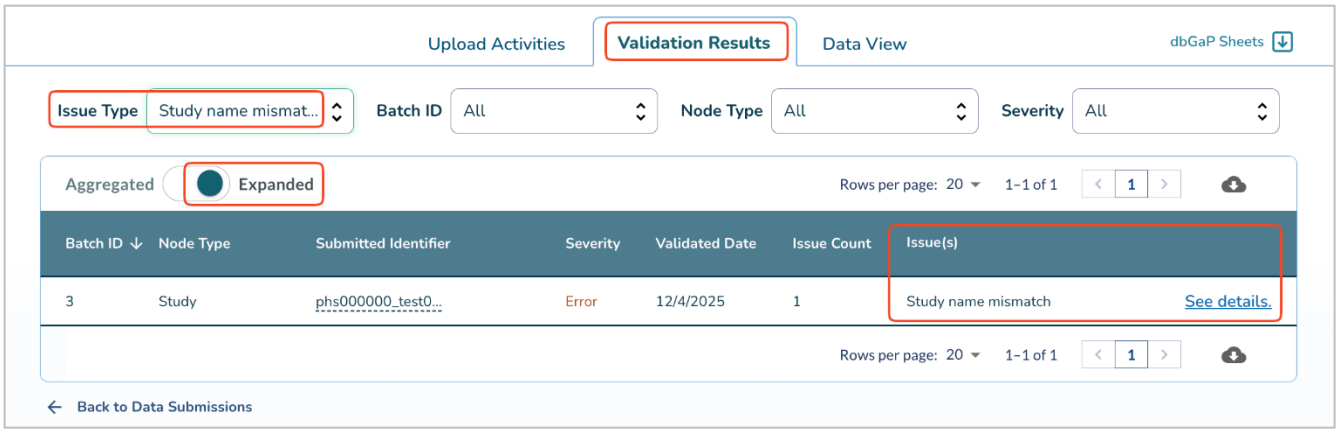

Figure 33. Expanded Validation Results Table

Validation Table Features: This table shows all the errors that were found after the validations were run. The information in the columns can be interpreted as follows:

- **Batch ID** – This correlates with the Batch ID shown on the Data Activity tab and indicates the specific upload with which the error is associated. This helps to identify which files may be involved.
- **Node Type** – These correspond to the various metadata submission templates. In the example above, selecting ALL in the Node Type displays all nodes with errors. The figure demonstrates that the error is located within the Demographic metadata template.
- **Submitted Identifier** – This is the identifier supplied by the user and is not a CRDC Submission Portal identifier. Again, this should specifically identify what object is causing the error.
- **Severity** – Severity will either be *Error* (which must be corrected before the submission can be finalized) or *Warning* (which should be fixed but is not required to be fixed) .
- **Validated Date** – This is the date that the validation was run.
- **Issues** – This gives a brief description of the error and a link to open a dialog box with more details about the error. Figure 34 presents an example of validation issue details.

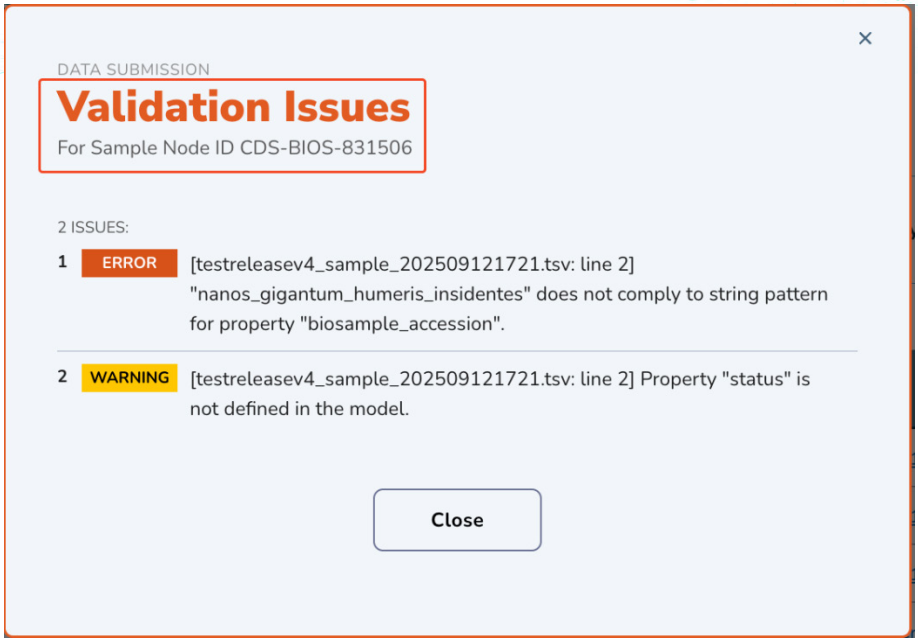

Figure 34. Validation Error Details

4.1.2. Requesting Permissible Values (PV)

Automatic validations in the Submission Portal are performed primarily to abide by the rules defined by the selected Data Model for data submission. In addition, the Cancer Research Data Commons (CRDC) requires that validations on data be performed against **Common Data Elements (CDEs)** and **Permissible Values (PVs)** as defined by the **caDSR standard.**

During the validation process, if an error is flagged under the Issue Type **value not permitted** in the Validation Results Table, a **Request for a New Permissible Value (PV)** must be submitted by the Submitter. See Figure 35 and 36. The New PV Requested should identify the desired value that best represents the data, along with a justification for its use.

The system allows only one PV request per CDE. Decisions on PV requests will be reviewed, and the outcome will be communicated by the assigned **Data Concierge**.

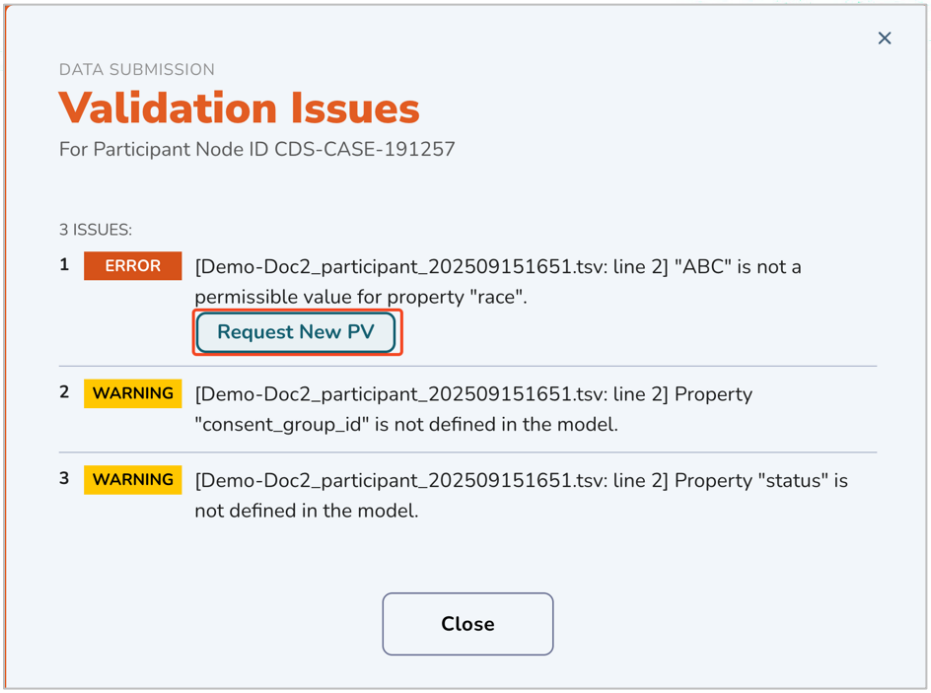

Figure 35. Request New PV

Figure 36. Request New PV Form

5.2 Revising Released Metadata – Options to Keep or Overwrite Existing Content

After uploading the metadata to the CRDC Submission Portal and performing the validation process, if your metadata files include properties with values that already exist in the system from a previously released data submission, the system will generate a warning message. You can view these warnings in the Validation Results table by selecting the issue type - **update existing data**. Figure 37 shows how the warning messages appear for this Issue Type and Node Type *Genomic_info*, and Figure 38 illustrates the detailed information displayed when you click the "see details" link in the **Issues** column.

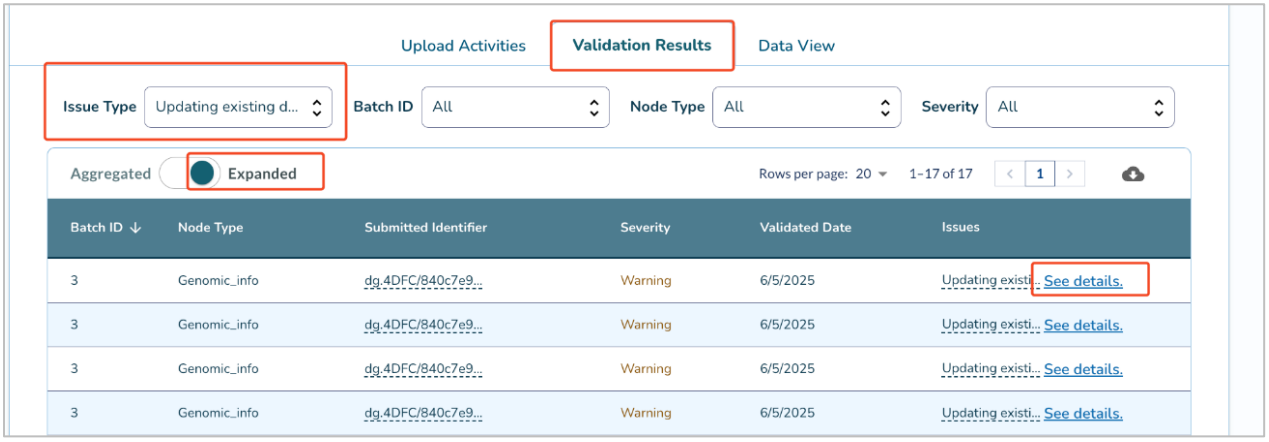

Figure 37. Visualize Validation Issue Type – Update Existing Data

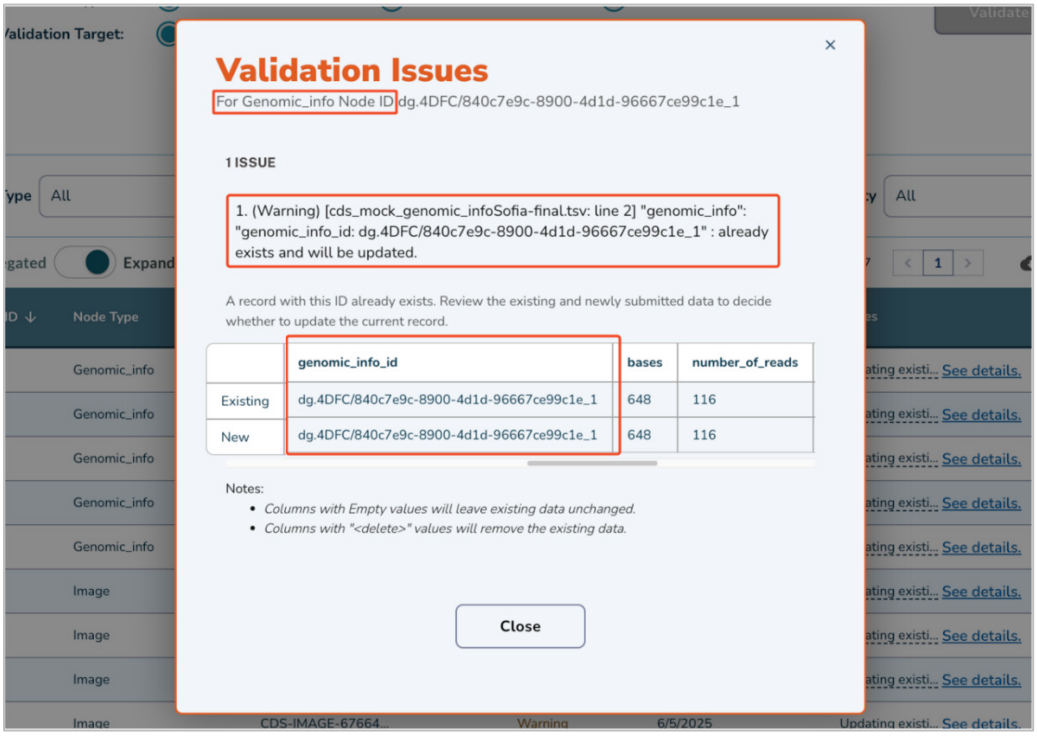

Figure 38. Illustration of the Details of Warning Message – Update Existing Data

It is important to follow the guidelines below to ensure the existing information from previous submissions is not unintentionally removed or overwritten. Note that validations will not be complete until the necessary actions are taken.

What Happens Automatically

- If a property or column in your metadata file (tsv) is left blank or not included, the system will keep the existing value for that field. See Table 1.
    - This applies to both required and optional fields.
    - For example, if the existing data for a sample has "sample_type = Tumor," and you leave sample_type blank or omit this property in your metadata file, the system will keep "Tumor."

What Submitter Should Not Do

- **Do not** leave the fields **blank** if your intention is to delete or clear a value. Blanks will not remove data.
- Do not try to remove required fields using special commands (see below); required fields cannot be deleted.

How to Delete a Value (if needed)

- If you need to **delete** a value from the previously released submission (e.g., an incorrect entry), use the command word <**delete>** as the value for that specific property.
- It is not **case-sensitive**: <delete>, <Delete>, and <DELETE> all work the same. This word must appear exactly as shown, with angle brackets (< >).
- Use this only for **optional fields**. If you try to delete a **required field**, your data submission cannot be submitted.

Property Type	Property Name	Existing Value in System	Value in Your Metadata file	What Happens?
Required	sample_type	Tumor	*(blank)*	Keeps "Tumor"
Required	sample_type	Tumor	<delete>	❌ Error – can't delete required field
Optional	sample_type_category	DNA	<delete>	Deletes DNA
Optional	sample_type_category	DNA	*(blank)*	Keeps DNA
Optional	sample_type_category	*(empty)*	*(not included)*	Still empty, nothing changes

Table 1. Summary of how existing metadata is updated during submission

5.3 Correcting Errors

Errors should be corrected by addressing the issues in local files, re-uploading the corrected file, and running the validation again. This process should be repeated until all errors have been addressed, and the validation returns no errors.

Anything marked as an *Error* in the Severity table must be fixed before the dataset can be formally submitted. Anything marked as a *Warning* will not block the final submission; however, users are ***strongly encouraged*** to fix warnings as well.

5.4 Remove Specific Files

After validating the uploaded data and metadata files, users can view the high-level details of these files in the **Data View** table. As shown in Figure 39, select either **data file** or **file** from the Node Type dropdown menu. To remove a specific data file or file ID along with its associated metadata, select the file and then click the **delete** icon. Users can also download the contents displayed in the Data View table as a TSV file by choosing the Node Type and clicking the **download** icon.

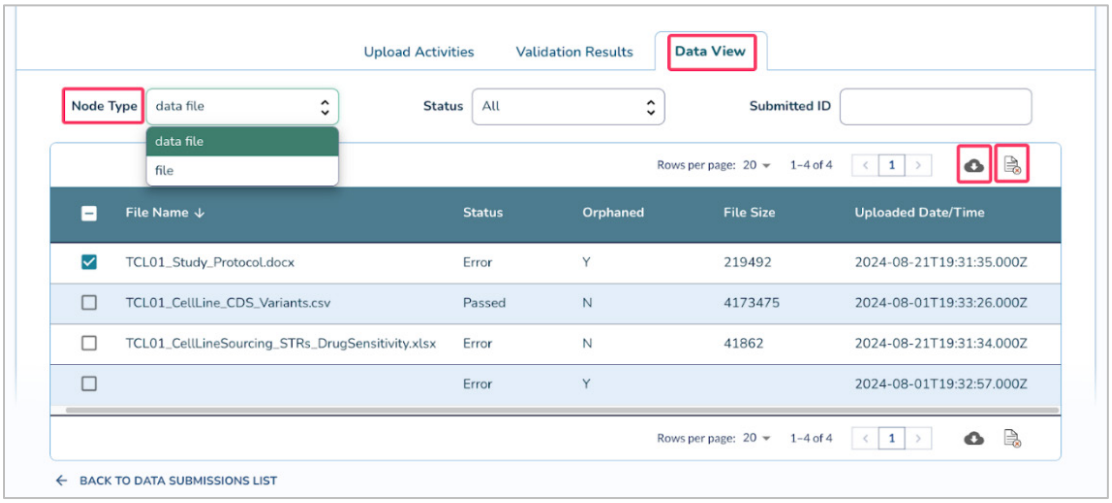

Figure 39. Removing Specific Files

The contents of the **Data View** table can be interpreted as:

- **File Name** - This column displays the name of the uploaded data file.
- **Status** - This indicates the validation status, which can be *Error*, *Warning*, or *Passed*. Details of any errors are available under the Validation Results. Warnings do not halt the submission process.
- **Orphaned** - This column shows whether the data file has associated metadata uploaded in the system. *Y* means the file is orphaned and lacks associated metadata, while *N* means the file has associated metadata uploaded in the system.
- **File Size** - This column lists the file size in bytes.
- **Uploaded Data/Time** - This column records the date and time when the file was uploaded.

6. Submitting Your Final Dataset

When a dataset has passed all validations with no outstanding errors, the Submit button at the bottom of the page is activated. Clicking the Submit button locks the submission and passes control to the Data Concierge for a final check. No further changes will be allowed. Should the Submit button be clicked in error, contact the assigned Data Concierge and they can reject the submission and return the control to you.

7. What to Expect After Submission

Once the final dataset has been submitted, the CRDC Submission Team will perform some final checks to make sure everything is as required by the destination Data Commons (for example, GC, ICDC, and CTDC). If those checks pass, the submission will be released to the appropriate CRDC Data Commons, and you will receive notification that the Data Commons is now responsible for the next steps. The respective Data Commons will be responsible for indexing and releasing the files for secondary sharing and will be made available and accessible on their portal.

If the final checks reveal some unexpected issues or the Data Commons team has any questions or concerns, the Data Concierge for your submission will reach out with additional questions and may reopen the submission to allow additional corrections.