RADx-UP CDCC Non-CDE File

Submission Guidance

Updated March 22 2023

Table of Contents

Part One: Introduction	3
myRADx-UPhome	3
Data Managers	3
File Type	3
File Name	3
File Scope	4
Data Sharing Concerns	4
Participant Identifiers	4
Record_id	4
Upload Data Inputs	4
Appendix A: Data Dictionaries	7
Appendix B: Summary Statistics	11
Categorical CDEs	11
Continuous Variables	1
Appendix C: Qualitative Codebook Guidance	11
Codebook structure:	12
Codebook Format:	12
Example of a code definition:	12
Example Codebook Structure	14
Appendix D: RADx-UP Guidance for De-identifying Qualitative Transcripts*	15
Background:	15
De-Identifying Procedures:	15
Appendix E: Further Details	17
How to Submit Data	17
Submission Statuses	19





Part One: Introduction

This document is intended to provide guidance to RADx-UP projects on uploading study data that has not been mapped to CDE data such as qualitative data, EHR data, or HIE data. The data transfer process is designed to be as straightforward as possible. NIH strongly encourages the sharing of qualitative data so that this information may be broadly used and learned from.

If your project has non CDE files you would like to upload, please contact your Engagement Impact Team (EIT) to facilitate the process.

You can use the following links to jump to the appropriate sections:

- <u>Data Dictionaries</u>
- Summary Statistics
- Qualitative Codebooks
- <u>De-Identifying Transcripts</u>
- More Details on Submitting Data

MYRADX-UPHOME

myRADx-UPhome is a secure web portal that all RADx-UP projects have access to. You will be able to upload study data to the CDCC through this portal. Please do not use email or other means to share any participant-level study data with the CDCC.

Data Managers

A few members of your project team, chosen by the project principal investigator (PI), will be given access to upload study data via *my*RADx-UP*home*. These users will also receive data upload notifications, reminders, and associated feedback. Please contact your EIT to make changes to the list of users who can upload data from your project.

You need specific permission to access the upload page for non-CDE files. This is a different permission than CDE upload permission assigned to your data manager.

Please contact your EIT to make changes to the list of users who can upload non-CDE files from your project.

FILE TYPE

Projects will NOT be allowed to upload an executable file, images, zip files, or videos. Other file types will be accepted. You must review data sharing concerns before uploading files, the CDCC does not check that your Data Transfer Agreement (DTA) matches the data uploaded.

FILE NAME

Each file you upload with a *unique name* will be added as a separate file to our database. If you wish to update or add additional data to a currently uploaded file, name the file with the exact same name, and it will replace the current file within the database. You can also delete or download files from the File History section.



FILE SCOPE

You may upload stand-alone data or linked to the CDE data that you are submitting to the CDCC. If participants overlap between CDE data and non-CDE data, they should share a common record ID.

For linked data such as quantitative data beyond the CDEs or linked data to be linked to CDEs, it is vital to make sure they share a participant record ID.

DATA SHARING CONCERNS

Please upload only the data elements that are covered under a participant's consent. If certain participants did not consent to share their data with the CDCC, please exclude their data from the study data files uploaded to *my*RADx-UP*home*. If your study design allows optional sharing of additional identifying information with the CDCC, please design your process so that the optional shared data matches the individual consent.

If you are anonymizing your participant identifiers, please ensure that the identifiers remain consistent between data uploads. This helps us identify any potential quality and data integrity issues over time.

If you're using MRN as record ID but are not consented to share identifiable data with CDCC, please change the record IDs to anonymous IDs before submitting. Please refer to the <u>Health and Human Services de-identification quidelines</u>.

DTA security level	Data NOT Allowed
PHI	all data allowed
Limited	Names, Geolocation smaller than state, dates (not years), age over 90, Telephone / Fax numbers, email, SSN, Medical record numbers, Health plan beneficiary numbers.
Deidentified	"Limited" restrictions above + city, any dates

PARTICIPANT IDENTIFIERS

record_id: record_id can be anonymized but the de-identified id should be consistent across data uploads for a given participant.

If participants overlap between CDE data and non-CDE data, if possible we recommend they share a common record ID.

UPLOAD DATA INPUTS

When submitting a file, you will be required to provide the following information about the data you are submitting. Please ensure that this information is accurate as it will be critical in the usability of this data.

- **Short Description:** Enter a short description of the file which will be displayed along with the file. (Maximum of 255 characters).
- <u>Participant Identifiers:</u> Are participant identifiers (record_ids) consistent between the non-CDE file being uploaded and the CDE file you have already uploaded?
 - o Options:



- Matches CDE File: The file contains record_ids that can be linked to the record_ids in the CDE upload file. If possible, this is the preferred option so your data can be properly linked.
- Does NOT match CDE File: The file contains record_id's but some or all of them do not match the participants who are in your CDE uploads.
- **Unknown if file matches CDE Files:** You do not know if there are record_ids that may or may not match.
- NA no participant identifiers: There are no record_ids in the file. This applies to data dictionaries and summary statistics.
- <u>Privacy:</u> What is the privacy level of the data contained in the file you are uploading? Privacy levels are specified on the <u>Health and Human Services de-identification guidelines</u>. (more specifics shown in Data Sharing Concerns section below).
 - o Options:
 - PHI
 - Limited Data Set
 - De-identified Data
 - NA no participant identifiers
- File Type: What type of file are you uploading?
 - Options:
 - EHR
 - HIE
 - Qualitative
 - Data Dictionary
 - Summary Statistics
 - Other

There are some fields you can choose to provide:

- More Details: Enter any pertinent details about the file you are uploading (Maximum of ~10 pages).
- <u>File Type Format:</u> If you are uploading Summary Statistics or a Data Dictionary, please specify the format. All of these format types can be found <u>here</u> for download.
 - Options:
 - Data Dictionary-row per AnswerCde csv-3.00: The standardized data dictionary format provided by the CDCC that you may fill out. Each row corresponds to an answer to a CDE question. When uploaded, you upload as a pipe delimited file, which will have a .csv filename extension.
 - Data Dictionary-redcap csv-2022.oo: The basic REDCap data dictionary export saved as a csv.
 - Data Dictionary-redcap xlm-2022.00: The basic REDCap data dictionary export saved as an xml.
 - Summary Statistics-continuous csv-1.00: The standardized summary statistics format provided by the CDCC for continuous CDEs.
 - Summary Statistics-categorical csv-1.00: The standardized summary statistics format provided by the CDCC for categorical CDEs.



- Data Documentation Initiative (DDI): standardized format for behavioral, social and economic sciences. More information available here.
- Fast Healthcare Interoperability Resources (FHIR): standardized format for healthcare data. More information available here.
- NA-not a recognized format version type.



Appendix A: Data Dictionaries

Before project closeout, your team will need to send a CDE data dictionary to the CDCC to store for future use. This will facilitate understanding of your dataset, especially if you made any changes to the CDEs by changing answer options or wording or if your file contains non-CDE questions. If you are including non-CDE questions in your CDE file upload, the data dictionary should describe these questions as well. There are two options to achieve this, detailed below. You can create your own data dictionary, or the CDCC can create the standardized data dictionary format for you. Your team can modify this format to adhere to changes you made to approved questions.

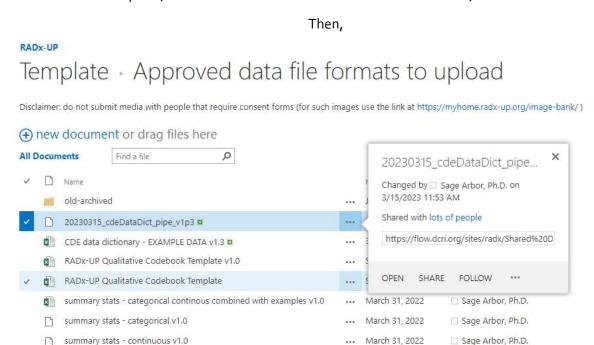
PROJECT UNIQUE DATA DICTIONARY FORMAT

You can email your EIT and informaticist a REDCap data dictionary (downloadable from your REDCap project), but we ask that you add the "questionSource" column. If you are not using REDCap, you can find the approved format for the data dictionary here and an example of how it looks filled out.

APPROVED STANDARD DATA DICTIONARY FORMAT

The approved data dictionary format allows CDE metadata to be queried by future researchers across all projects. If you plan to upload your data dictionary using this format, please contact your informaticist. Your informaticist will create a custom data dictionary document for you based on CDEs you have already uploaded. Your team can modify this format to accommodate CDE changes (the stem or answers) or added survey guestions.

<u>This Sharepoint folder</u> contains sample data for various approved data file formats. To download any file from this Sharepoint, click the three horizontal dots next to the file itself, as shown here.



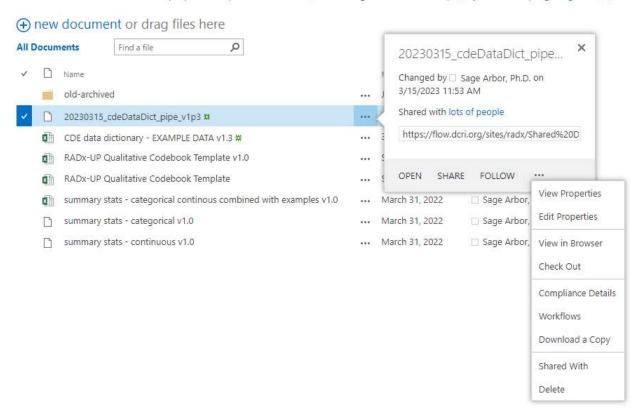
Then, click on the three dots once more and click "Download a Copy", as shown on the next page.



RADx-UP

Template - Approved data file formats to upload

Disclaimer: do not submit media with people that require consent forms (for such images use the link at https://myhome.radx-up.org/image-bank/)



In particular, this Sharepoint folder contains an example data dictionary spreadsheet (CDE data dictionary – EXAMPLE DATA v.1.30). In addition, it contains a data dictionary spreadsheet (20230315_cdeDataDict_pipe_v1p3) with the expected pipe-delimited format.

The columns in this standard are:

- questionVariable: the question text associated with a CDE variable
- answerVariable: the variable name of the CDE
- answerValue: the numerical value associated with an answer option for this CDE
- answerText: the text associated with the numerical value
- **showOnlyIf**: the Branching Logic used to show
- <u>questionType</u>: the type of question; radio refers to a drop-down of options, checkbox refers to a multiselect, text refers to a free text field, date refers to a date entry.
- <u>answerType</u>: the format of the answer (int for integer, text, date, varchar, etc.)
- questionSource: Please list data sources for all of your variables / data elements.
 - o If a measure comes from a validates scale, please list all identifying information for that scale such as name and version number. For example, if you are using a PROMIS Pediatric Anxiety scale which has multiple versions, do not list as PROMIS Pediatric Anxiety but list as: PROMIS Pediatric Short Form v2.0 Anxiety 8a



- o Examples of data sources:
 - RADx-UP CDE (radxup)
 - Self developed by your project team (selfDeveloped)
 - Phenotypes and eXposures Toolkit (phenx)
 - Patient-Reported Outcomes Measurement Information System (PROMIS)
 - Behavioral Risk Factor Surveillance System (BRFSS)
 - Mini-Mental State Exam (MMSE)
 - Quality of Life in Neurological Disorders (NeuroQol)
 - Patient Health Questionnaire (PHQ9)
 - State-Trait Anxiety Inventory (STAI)

See the below example of a radio/drop-down variable, pregnancy_status.

questionCDE		answerCDE	answerVal answerText	showOnlyIf	questionT	ype answerTy	pe questionSource
pregnancy_status	Are you currently pregnant?	pregnancy_status	0 Not Pregnant	(bio_sex_birth = 1) OR (bio_sex_birth_2 = 1)	radio	int	radxup
pregnancy_status	Are you currently pregnant?	pregnancy_status	1 Pregnant	(bio_sex_birth = 1) OR (bio_sex_birth_2 = 1)	radio	int	radxup
pregnancy_status	Are you currently pregnant?	pregnancy_status	98 Don't know	(bio_sex_birth = 1) OR (bio_sex_birth_2 = 1)	radio	int	radxup
pregnancy_status	Are you currently pregnant?	pregnancy_status	99 Prefer not to answer	(bio_sex_birth = 1) OR (bio_sex_birth_2 = 1)	radio	int	radxup

See below example of a check box variable, race_ethn_race.

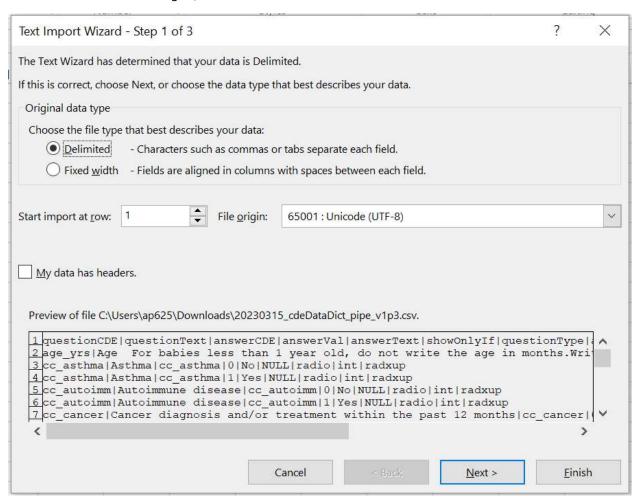
questionCDE		answerCDE	answerVa	l answerText	showOnlyIf	questionTyp	e answerTyp	e questionSource
race_ethn_race	What is your race?Mark one or more	race_ethn_race	1	O American Indian or Alaska	NULL	checkbox	bit	radxup
race_ethn_race	What is your race?Mark one or more	race_ethn_race	1	1 American Indian or Alaska	NULL	checkbox	bit	radxup
race_ethn_race	What is your race? Mark one or more	race_ethn_race	_1!	O Some other race	NULL	checkbox	bit	radxup
race_ethn_race	What is your race?Mark one or more	race_ethn_race	1!	1 Some other race	NULL	checkbox	bit	radxup
race_ethn_race	What is your race?Mark one or more	race_ethn_race	2	O Black or African American	NULL	checkbox	bit	radxup
race_ethn_race	What is your race?Mark one or more	race_ethn_race	_2	1 Black or African American	NULL	checkbox	bit	radxup
race_ethn_race	What is your race? Mark one or more	race_ethn_race	_3	0 Asian	NULL	checkbox	bit	radxup
race_ethn_race	What is your race?Mark one or more	race_ethn_race	3	1 Asian	NULL	checkbox	bit	radxup
race_ethn_race	What is your race?Mark one or more	race_ethn_race	4	Native Hawaiian or Other	NULL	checkbox	bit	radxup
race_ethn_race	What is your race?Mark one or more	race_ethn_race	_4	1 Native Hawaiian or Other	NULL	checkbox	bit	radxup
race_ethn_race	What is your race?Mark one or more	race_ethn_race	5	0 White	NULL	checkbox	bit	radxup
race_ethn_race	What is your race?Mark one or more	race_ethn_race	_5	1 White	NULL	checkbox	bit	radxup



Please note that your data dictionary should be uploaded in a pipe delimited format with a .csv filename extension. To open the sample pipe delimited file in Excel, create a new spreadsheet before opening the pipe delimited file within Excel, as shown below.

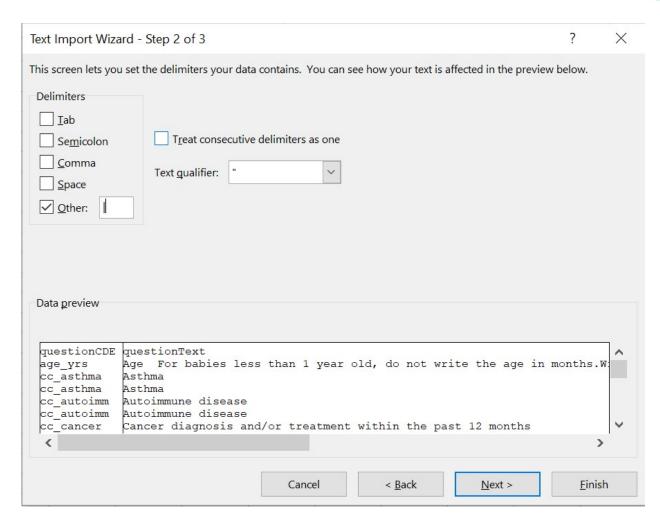


Then, the Text Import Wizard will appear. The first page will have file type marked as "Delimited" and should not need to be changed, as shown below.



For the next page, deselect "Tab" as the delimiter and check "Other", typing in the "|" pipe delimiter as shown on the next page. Then, click "Finish" to open your file as pipe delimited.





To ensure your Excel file is saved as pipe delimited, please refer to this quide.

Appendix B: Summary Statistics

Your team may be required to upload summary statistics. If you are, you will be notified. Regardless of requirement, you are welcome to upload summary statistics.

If you are uploading aggregated summary statistics to the CDCC, please reference the examples and use the approved format found here. This allows us to ingest your data and use it in summary numbers.

Appendix C: Qualitative Codebook Guidance

Much of the work in the RADx-UP initiative is done through qualitative methods either for primary outcomes or for preparatory phases of research. This qualitative data generates rich information about the needs and perspectives of underserved communities this initiative hopes to serve. NIH strongly encourages the sharing of qualitative data so that this information may be broadly used and learned from.

This guidance provides the framework for creating consistent codebooks for qualitative data across the RADx-UP consortium so that qualitative data that is shared may be well understood.



CODEBOOK STRUCTURE: Use columns in your codebook to indicate when you have top-level and sub-codes. Top-level codes are the highest-level codes with broad definitions. Sub-codes are subordinate or lower level codes with more specificity. A top-level code is not required to have a sub-code. However, sub-codes must have a top-level code by definition and may also have their own sub-codes.

Other terms you may have heard for coding hierarchies are parent and child codes, structural and thematic codes, etc.

Example:

When coding for "Testing Barriers" in a COVID-19 access to testing study, you may code
"Testing Barriers" as your top-level code and "Lack of Trust", "Access to Transportation", "Test
Availability", or "Misinformation" as sub-codes.

Codebook Format:

- Code name (columns are labeled as Top-level code and Sub-code in Excel) 1-2 words to describe the code name
- **Brief definition:** Approximately one sentence to briefly describe the concept.
- **Full definition:** Complete definition of the code that defines the concept and any relevant nuances. May include key words from the data that make it easier to identify.
- **Exemplar Quote:** A section of text that illustrates when the code should be applied.
- When not to use: May also be referred to as counter examples. Used for differentiation from conceptually similar codes. When necessary, provide an explanation and example of when the code should not be used and, if applicable, what alternative codes might be used.
- Origins: Is the code Deductive or Inductive?
 - O Deductive codes are developed prior to data analysis based on literature reviews, prior knowledge of the topic being studied, aims of the study, etc. These are often referred to as *α priori* codes.
 - o Inductive codes are codes that were not predetermined but arise during data analysis. Inductive codes are also referred to as emergent codes.
- Some studies may use one type of code or a mixture of both. Indicating the origin of each of your codes provides insight into how your codebook was developed.

EXAMPLE OF A CODE DEFINITION:

Top-Level Code: TESTING BARRIERS

Brief Definition: Barriers to COVID-19 Testing

Full Definition: An obstacle that prevents a person's access or willingness to get COVID-19 testing. Obstacles may be financial, geographic, institutional, relational, education related, etc.

Example: "I don't know where to get a COVID-19 test that is close enough to my work that I can get it on my break. My position doesn't have vacation or sick leave and it's hard to get to a testing center in the evenings and weekends with my kids. Besides that, I heard that the tests were not telling people if



they really have COVID so it probably would not have been useful and my symptoms were not that bad when I was sick."

When Not to Use: Do not use this code when a participant refers to barriers to anything other than COVID-19 Test access or willingness.

Origins: Deductive

EXAMPLE CODEBOOK STRUCTURE

Top-Level Codes	Level 1 Sub- code* [if applicable]	Level 2 Sub- code* [if applicable]	Brief Definition	Full Definition	Exemplar Quote	When not to use	Origins [D = Deductive, I = Inductive]
Testing Barriers							
	Lack of Trust						
	Access to						
	Transportation						
	Test						
	Availability						
	Misinformation						

^{*}If you have more than one level of child codes, insert a new column to the right of your last sub-code for each necessary level.

References:

MacQueen, K. A., McLellan, E., Kay, K., & Milstein, B. (1998). Codebook development for team-based qualitative analysis. CAM Journal 10(2):31-36. https://doi.org/10.1177/1525822X980100020301

Mihas, P., & (2019). Learn to build a codebook for a generic qualitative study. In SAGE Research Methods Datasets Part 2. SAGE Publications, Ltd. https://dx.doi.org/10.4135/9781526496058

Appendix D: RADx-UP Guidance for De-identifying Qualitative Transcripts*

BACKGROUND:

HIPAA (The Health Insurance Portability and Accountability Act of 1996) is a federal law that applies to information retrieved from medical records, medical providers, or for research such as RADx-UP. Thus, any time personal health information (PHI) is obtained from a participant's medical record, focus group, interview, or other data collection method and shared with the RADx-UP CDCC, the Privacy Rule (HIPAA) standards for de-identifying that information must be used, which means removing names, dates, locations, etc. (The Duke Health HIPAA policies can be found here: https://irb.duhs.duke.edu/policies-and-regulations/policies) *Make sure that you are following the current guidelines and policies of your project's IRB.

Additionally, because the RADx-UP CDCC is a large scale data coordinating center taking in data from over 130 projects, the CDCC uses Safe Harbor rules for assessing what data qualifies as a HIPAA identifier. This means that rather than using expert determination to say whether a data element is a HIPAA identifier based on the data element definition, related data, and study design, we assume that all data elements falling into a PHI category are PHI. For example, all dates are PHI because dates are a category of PHI. This allows us to standardize our data privacy intake and data quality checks across projects.

As such, <u>information obtained in a participant interview or focus group</u> needs to be de-identified to the extent needed to protect participant confidentiality. This typically means removing only names or details in the data that may make someone identifiable. De-identifying transcripts, therefore, has to follow the HIPAA standards. Any elements of qualitative data that fall into one of the listed HIPAA PHI categories, including names, dates, geographic areas, or other details that could be used to easily identify someone will need to be changed. Because there are no specific de-identifying guidelines to follow for rich, qualitative data, sometimes deciding how to de-identify data can be challenging. Thus, de-identifying data may require careful review of unique situations and case-by-case decisions.

The following procedures should be used by RADx-UP projects to de-identify qualitative data:

DE-IDENTIFYING PROCEDURES:

- 1. Create a copy of the original transcript, add "de-identified" to the document file name, and save the copy in a new, separate folder. Work from the copy.
- 2. Remove the study participants' names, first and last. Replace with a participant code (such as: Participant 1 or P1, etc.). *Note: The code cannot be the participant's initials or based on any other identifying information about the participant such as their date of birth or phone number.
- 3. Remove the names of participants' family members. Replace with XXXX. If necessary, add the person's role such as husband, daughter, son, etc. in parentheses.



Example: My husband, Michael, took me to every chemo appointment.

De-identified: My husband, XXXX, took me to every chemo appointment.

Example: "I said, Michael, I don't think I can handle doing three extra treatments."

De-identified: "I said, XXXX (husband's name), I don't think I can handle doing three extra treatments."

- 4. Remove references, smaller than State, to <u>specific</u> streets or addresses, including neighborhood, city, county, precinct, or zip code, where participants live. Replace with XXXX.
- 5. Remove all references to dates, except year; unless the participant is over the age of 89. If a person is 90 or above, remove all references to date of birth, including the year, and replace with YYYY, or "90 or above." Additionally, remove all references, including year, to any other date that may indicate that the person is 90 or above.

Example of a date that must be removed for a participant 90 and above: "Let's see... I was 65, so it was in 1995 when I had my last round of chemo". In this case, you must redact the reference to the year so the person's age cannot be ascertained by the knowledge that they were 65 in 1995.

Example of a date that is acceptable to leave in for a participant 90 and above: "Let's see... I had my last round of chemo in 1995." In this instance, the information does not tell us how old the person reporting is, therefore it is fine to leave the date reference in the transcript.

- 6. Any other data that might identify the participant, such as their employer, profession, or unique stories should be reviewed on a case-by-case basis and changed as needed.
- 7. Names of doctors, nurses, hospitals, and other members of the care team should be removed and replaced with XXXX. If necessary, add the healthcare provider's or entities role such as doctor, nurse practitioner, physical therapist, clinic, hospital, public health department, etc. **See example provided above in #3.
- 8. Think of de-identifying this way: If a random person read a transcript, would they be able to identify the participant based on the information in the transcript?

*Guidance for this document was provided by the qualitative research team at Northwestern University, Feinberg School of Medicine, Department of Medical Social Sciences.

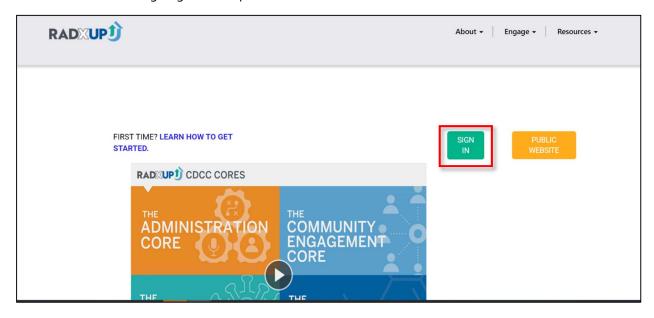


Appendix E: Further Details

Note: The screenshots in this section may be out of date, but the information will remain the same. This section was last updated 10/20/2021.

HOW TO SUBMIT DATA

1. Go to myhome.radx-up.org and click Sign In. Sign in through your InCommon Account or see the section "Signing in with Sponsored Accounts".

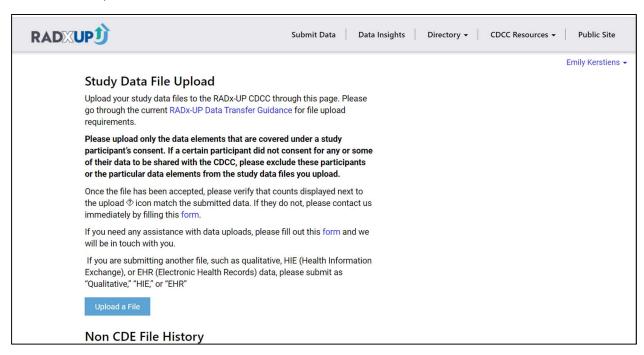


2. Under the "Data Management" tab, select "Data Submission Non CDE"

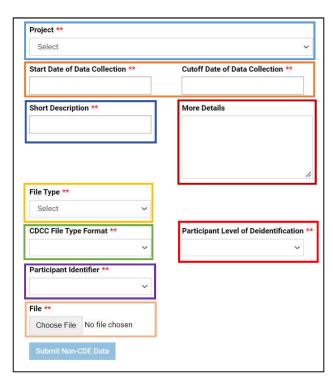




3. Click 'Upload a File'



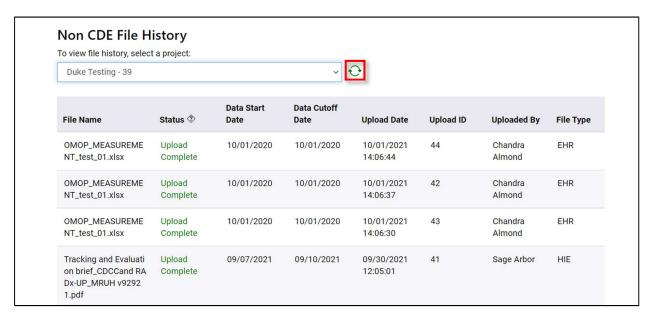
- 4. Select your project from the drop-down menu (most people will only have one project)
- Add start date and cutoff date for the file you are submitting. Remember the file is cumulative and the start date should be the same for every upload.
- 6. Enter a short description (less than 255 characters) of the file.
- 7. If you'd like, enter more details to provide additional context.
- 8. Select File Type
- 9. Select a File Type Format. If your file
- 10. Select if the Participant Record IDs match in the Participant Identifier box.
- 11. Select the level of de-identification in this file.
- 12. Select your file.





13. Submit the file!

After this, your file will appear in the Non CDE File History. Click the green "Refresh" • to view progress as the status changes. Once it says "Upload Complete", the process is done.



SUBMISSION STATUSES

Once a file is uploaded, there are a number of statuses your file can take on. After clicking "Submit", the job will be shown in the job log table at the bottom of the page with the temporary status "Initiated" as the file processes. If you click the fresh arrow \bigcirc , you will see an updated status.

- "Rejected": If the file is bigger than max file size, it will be rejected.
- "Upload Complete": the file has uploaded to the CDCC.

A summary of the statuses can be seen in the following table.

Status	Description
Initiated	The file has been uploaded, processing has begun.
Upload Complete	The job execution has completed successfully.
Rejected	File rejected if it is bigger than the max file size.