

This document describes the types of data files, their naming conventions, and content rules when preparing data files to be submitted to the RADx Data Hub.

preorigcopy

This file represents a data file that was submitted to a (C)DCC for ingestion into the RADx Data Hub. An origcopy file that is derived from this preorigcopy file that satisfies the rules below should be submitted to the Data Hub. Investigators and/or (C)DCCs should retain this preorigcopy file.

origcopy

origcopy files represent files that contain data that are ready for analysis by Data Hub users but that have not been harmonized (i.e., Tier 1 Common Data Elements (CDEs)).

- MUST have a file suffix of `_DATA_origcopy.csv`
- MUST only contain de-identified data
- MUST be UTF-8 encoded
- MUST be a well-formed Comma Separated Values (CSV) file; each row/record in the CSV file MUST have the same number of columns
- MUST have a single non-empty header row
- MUST NOT have any completely empty rows/records
- MUST NOT have any completely empty columns
- MUST have a corresponding RADx metadata (META) file
- MUST have a corresponding RADx data dictionary (DICT) file
- MAY contain blank values (or empty cell values) in otherwise non-empty rows or columns
- SHOULD contain missing value codes in place of blank values (these are not required in origcopy files)
- MAY contain data elements that are Tier 1 harmonizable according to the rules in the RADx Global Codebook
- MAY contain data elements that are NOT Tier 1 harmonizable according to the rules in the RADx Global Codebook

Note: For end-users of the RADx Data Hub, origcopy files will be known as non-harmonized files.

transformcopy

transformcopy files represent files that contain data that have been harmonized with respect to the harmonization rules in the RADx Global Codebook. They are derived from origcopy files by applying said harmonization rules and by replacing blank values with RADx Missing Value codes.

The sole difference between an origcopy file and a transformcopy file is Tier 1 data harmonization.

Any origcopy file that contains data elements that are harmonizable to Tier 1 CDEs MUST have a corresponding transformcopy file that:

- MUST have a file suffix of `_DATA_transformcopy.csv`
- MUST only contain de-identified data
- MUST be derived from an origcopy file

- MUST contain ALL data elements that are contained in the origcopy file
- MUST contain harmonized data for ALL Tier 1 CDEs for which there are NO approved exceptions; Tier 1 CDEs for which there are approved exceptions do not need to be included in the transformcopy file.
- MUST NOT contain ANY blank values. All blank values MUST be replaced with RADx Missing Value codes. This applies to ALL data copied over from the origcopy file in addition to any harmonized data.
- MUST be UTF-8 encoded
- MUST be a well-formed Comma Separated Values (CSV) file; each row/record in the CSV file MUST have the same number of columns
- MUST have a single non-empty header row
- MUST NOT have any completely empty rows/records
- MUST NOT have any completely empty columns
- MUST have a corresponding RADx metadata (META) file
- MUST have a corresponding RADx data dictionary (DICT) file
- MUST be harmonized with respect to the harmonization rules that are defined in the RADx Global Codebook. All harmonizable data elements in the source origcopy file MUST be mapped to Tier 1 CDEs in the RADx Global Codebook.
- MAY contain data elements that are NOT Tier 1 harmonizable according to the rules in the RADx Global Codebook

Note: For end-users of the RADx Data Hub, transformcopy files will be known as harmonized files.