

DOCUMENT REF.: Not Applicable	LAST REVISION DATE: N/A	DOCUMENT TYPE: Not Applicable		VERSION NO.: v1.0
	DOCUMENT TITLE: SOP for Internal Data Organization and Validation	SPONSOR: National Institute of Mental Health (NIMH)	PROJECT NAME and / or PROTOCOL NUMBER: NeuroHIV Data Ecosystem	VERSION DATE: 30JUL2026
				EFFECTIVE DATE: 05AUG2026



SOP for Internal Data Organization and Validation

For

The National NeuroHIV Tissue Consortium

NeuroHIV Data Ecosystem

Sponsored by:

National Institute of Mental Health (NIMH)

DOCUMENT REF.: Not Applicable	LAST REVISION DATE: N/A	DOCUMENT TYPE: Not Applicable		VERSION NO.: v1.0
	DOCUMENT TITLE: SOP for Internal Data Organization and Validation	SPONSOR: National Institute of Mental Health (NIMH)	PROJECT NAME and / or PROTOCOL NUMBER: NeuroHIV Data Ecosystem	VERSION DATE: 30JUL2026
				EFFECTIVE DATE: 05AUG2026

TABLE OF CONTENTS

1.	PURPOSE/SCOPE	3
2.	DATA INGESTION	3
2.1.	Ingestion.....	3
3.	DATA VALIDATION	3
3.1.	Validation Checks	3
4.	DATA INTEGRATION/ORGANIZATION	3
4.1.	Data Flow for Internal Data Organization and Validation.....	3
5.	DATA LINKAGE	4
5.1.	Information Collected if Investigators Submitted to Other Repositories	4
6.	ABBREVIATIONS AND DEFINITIONS	4
7.	DOCUMENT HISTORY	4

DOCUMENT REF.: Not Applicable	LAST REVISION DATE: N/A	DOCUMENT TYPE: Not Applicable		VERSION NO.: v1.0
	DOCUMENT TITLE: SOP for Internal Data Organization and Validation	SPONSOR: National Institute of Mental Health (NIMH)	PROJECT NAME and / or PROTOCOL NUMBER: NeuroHIV Data Ecosystem	VERSION DATE: 30JUL2026
				EFFECTIVE DATE: 05AUG2026

1. PURPOSE/SCOPE

The purpose of this SOP is to cover procedures outlining the ingestion, validation, and organization of data housed in the repository.

2. DATA INGESTION

This section outlines procedures for receiving and organizing new datasets into the repository.

2.1. Ingestion

During ingestion, the ingestion tool processes datasets uploaded by investigators.

- The ingestion tool processes datasets one row at a time. Rows with errors will not cause the entire ingestion to fail.
- Errors encountered when writing a row to the corresponding database table are logged to AWS CloudWatch logs and a discrete log table (S3_PROCESSOR_LOGGER) within the database.

3. DATA VALIDATION

This section outlines methods for checking the accuracy, completeness, and consistency of incoming data.

3.1. Validation Checks

Several checks have been implemented to validate data during ingestion.

1. Empty file – Verify mapping and data file contain data before processing.
2. Null/Required – Ensure required (non-null) fields are populated.
3. Data type – Validate data types match those specified in the mapping template.
4. Length – Ensure field length matches specifications from mapping document. If mapping document does not specify, assign a default based on data type.
5. White space – Clean any white space in upload file.
6. Leading zero – Preserve leading zeros in fields where they are significant; ensure they are not lost during data processing by treating them as numbers.

4. DATA INTEGRATION/ORGANIZATION

This section outlines procedures for receiving and organizing new datasets into the repository.

4.1. Data Flow for Internal Data Organization and Validation

1. Lambda trigger invokes the ingestion function, one invocation per file.
2. Ingestion function parses in data file. Large files (>25MB) are broken up into smaller pieces and processed in parallel with the use of a queue (AWS Simple Queue Service [SQS]).
3. Ingestion function performs validation and handles errors as detailed in the During Ingestion section above. Validation checks are also listed above.
4. Ingestion function writes rows that have passed validation to the database.

DOCUMENT REF.: Not Applicable	LAST REVISION DATE: N/A	DOCUMENT TYPE: Not Applicable		VERSION NO.: v1.0
	DOCUMENT TITLE: SOP for Internal Data Organization and Validation	SPONSOR: National Institute of Mental Health (NIMH)	PROJECT NAME and / or PROTOCOL NUMBER: NeuroHIV Data Ecosystem	VERSION DATE: 30JUL2026
				EFFECTIVE DATE: 05AUG2026

5. DATA LINKAGE

This section lists information collected if investigators submitted their data to a linked repository. For example, if you have submitted sequencing data to dbGAP, you are **not** required to submit those data in duplicate to the neuroHIV Data Ecosystem. You **are** required to provide external repository information to enable the neuroHIV Data Ecosystem to provide links to any other repository where related data is available.

5.1. Information Collected if Investigators Submitted to Other Repositories

The Manifest File Metadata document completed for each data submission requires confirmation of the following identifiers for data submitted to linked repositories:

- Repository name
- Accession number/digital object identifier/collection ID
- A link to the dataset, if available

The NeuroHIV Data Ecosystem will establish hyperlinks to external related data.

6. ABBREVIATIONS AND DEFINITIONS

Abbreviation	Definition
AWS	Amazon Web Services
dbGAP	Database of Genotypes and Phenotypes
HIV	Human Immunodeficiency Virus
ID	Identifier
MB	Megabyte
SOP	Standard Operating Procedure
SQS	Simple Queue Service

7. DOCUMENT HISTORY

Version No.	Summary of Changes
1.0	Initial document