UT Health San Antonio
Joe R. & Teresa Lozano Long School of Medicine

Department of Population Health Sciences
Clinical Research Informatics Division

STANDARD OPERATING PROCEDURE

SOP Number: CRI.SOP. DMLC-003	**Title: Integration of External Data**		
Version No.: 0.0	**Effective Date: DRAFT**		**Page 1 of 11**
Supersedes Version: N/A **Dated:** N/A	REQUIRED APPROVALS BELOW		
CRI Director:	DocuSigned by: Meredith Zozus CE4632E6FE3443...	**Date:** 3/13/2024	
CISIL Approver 1:	DocuSigned by: Melanie Zuniga Rapp A01236D48E3A458...	**Date:** 4/3/2024	
CISIL Approver 2:	DocuSigned by: Becki Gerwitz 05121BFDB1724D1...	**Date:** 4/3/2024	

1.0 Purpose

This procedure specifies the process for the acquisition and integration of external data. The purpose of the procedure is to ensure that the acquisition of new data complies with regulations and other requirements and that consistent and appropriate procedures are followed for integrating external data with data managed by CRI.

2.0 Scope

This procedure applies to clinical studies managed by the Clinical Research Informatics Division (CRI), with external data other than those key entered into electronic CRFs, and to all CRI faculty, staff, and contract informatics employees performing data acquisition and integration tasks. This procedure covers system interfaces, data transfers, and establishing and maintaining linkage or referential integrity with data obtained from sources external to CRI or managed externally to CRI.

3.0 Responsibility

3.1 The CRI Directors will ensure that all personnel who perform data acquisition and integration tasks are trained on and comply with this procedure.

3.2 The Clinical Research Informatics Specialist (CRIS) shall identify all data on a project that originate or are managed outside the CRI Quality Management System (QMS), obtain and document project decisions which will be integrated with data managed by CRI, and complete data transfer and integration specifications (Attachment 2) for all data acquired and integrated by CRI.

3.3 UTHSA HOP 5.8.22 Data Protection outlines the requests requiring Patient Data Governance review and approval. The CRIS is responsible for ensuring appropriate approval before data acquisition by CRI. The CRIS may need to assist investigators or clinical leaders in completing a **D**ata **A**cquisition, **A**ccess, **U**se, and **R**elease (DAUR) form and submission of the DAUR Form to the UTHSA Patient Data Governance process.

3.4 CRI faculty and staff who perform data acquisition and integration tasks shall adhere to this approved SOP.

4.0 References

4.1 CRI.SOP.DMLC-001 *Data Management Plan Creation and Maintenance*

4.2 CRI.SOP.DMCL-004 *Data Collection and Processing*

4.3 HOP 5.8.4 Access Management

4.4 HOP 5.8.21 Data Classification

4.5 HOP 5.8.22 Data Protection

5.0 Definitions

Bulk Data: Multiple records of data about one or more individuals received at the same time, usually but only sometimes at an established recurring frequency. Bulk data may be identified or de-identified. In the case of data migration or receipt of legacy data or a data snapshot, bulk data may be received as a one-time data transfer.

Streamed data: Data about individuals or events pushed one at a time, in real-time or near real-time, and on an ongoing basis from another system through an established interface.

Messaging data: Data about individuals or events pushed or pulled one at a time, in real-time or near real-time, and on an ongoing basis from another system through an established interface where receipt confirmation is expected from the receiving system.

External data: Data that originate or are managed outside the CRI QMS.

Data integration: Data integration is the process of aligning data with study patients and study time points at which the data were obtained were collected.

Matching: The process of determining which data belong to the same entities, such as patients. Matching is usually required before record linkage.

Record linkage: Establishing a persistent association between data. The association serves as the mechanism through which the data are connected for reporting and analysis. Usually, record linkage is used to associate patient data from one source with data on the same individuals from another data source.

Deterministic record linkage: Using an identifier such as a patient number, study number, site number, visit number, or sample number to associate data. For example, CRF data and external lab data containing the patient identifier are “linked” by being labeled with that identifier. In deterministic record linkages, the value of the data elements used for the matching must be an exact match. Barring errors in the identifiers themselves, a deterministic match is an exact match.

Probabilistic record linkage: Inexact matching by one or more data elements is used when deterministic matching is not possible, or the error rate in the corresponding identifiers is high. For example, matching text strings, such as for last names that phonetically sound the same or are less than one or two characters different. An example of inexact matching using multiple data elements would be street address phonetically equivalent, last name phonetically equivalent, and two of either day, month, or year of the date of birth match. As inexact matches, the accuracy of probabilistic record linkages must be measured and reported.

6.0 Procedures

6.1 As part of data management planning, the CRIS identifies data for a study that originates outside the CRI QMS.

6.2 For each externally originated or managed data set, the CRIS, together with the study team and according to the Scope of Work, decides which will be integrated with data managed by or transferred and managed by CRI.

6.3 For each externally originated or managed data set, the CRIS, together with the study team, decides the attributes through which each data element should be associated with study entities such as but not limited to study sites, patients, time points, assessments, or biological samples.

6.4 For each externally originated or managed data set, other than the data key entered into eCRFs, the CRIS documents the following on the Data Integration Form (Attachment 1)

6.4.1 Whether the external data source is to be integrated by CRI or another member of the research team.

6.4.2 For data to be integrated by CRI, the CRIS documents the following:

6.4.2.1 Whether the data to be integrated are blinded and the names of the individuals who are unblinded.

6.4.2.2 Confirmation that the external data supplier is listed in the informed consent or HIPAA Authorization.

6.4.2.3 The timing of data transfer and integration, such as individual time points, or weekly, monthly, etc. and that the frequency and modality are consistent with CRIs and the external data provider's scope of work where applicable.

6.4.2.4 The mechanism through which data are received such as secure File Transfer Protocol (sFTP), encrypted email, or a system interface, and that the transfer mechanism is consistent with CRIs and the external data provider's scope of work where applicable.

6.4.2.5 The storage location/s of data received before integration.

6.4.2.6 The storage location/s of integrated data.

6.4.2.7 Whether an exchange or content standard will be used; if so the names and versions of the standards, or else data transfer specifications are required.

6.4.2.8 The party/ies responsible for checking the incoming data's consistency or other quality aspects. Note that the external data provider may run local quality checks and that integration checks are usually needed once CRI has matched the data; these checks may occur at multiple stages in the data processing.

6.4.2.8.1 How will exceptions be tracked to resolution

6.4.2.8.2 How exceptions will be communicated to the resolving party

6.4.2.8.3 Timeline within which resolutions are expected

6.4.2.9 The party responsible for investigating and resolving data problems, including making changes to the data and retaining the audit trail for such changes.

6.4.2.10 The data elements used to match incoming data to other study data

6.4.2.11 Whether the external data provider claims that the system and associated audit trail are 21 CFR Part 11 compliant.

6.4.2.12 The criteria for final acceptance of data from the external data provider and shall be cross-listed on the Database Lock checklist.

6.4.3 Changes to any items listed under the previous section except referenced data transfer specifications require revision and re-approval of the data integration form.

6.4.4 The Data Integration Form is a version-controlled document.

6.4.5 The study PI and Statistician or designee approve the Data Integration Form.

6.4.6 The Data Integration Form is stored in the DMP.

6.5 Data Transfer Specifications

6.5.1 The CRIS or designee drafts data transfer specifications that describe the format and content of transferred data in sufficient detail to support extract, transformation, and load programming required for data integration.

6.5.2 The CRIS is responsible for confirming that the data to be transferred are consistent with the protocol and target data dictionary.

6.5.3 The CRIS maintains version control with the external data provider on the data transfer specifications.

6.5.4 Data transfer specifications are stored in the Data Management Plan.

6.6 Transferred data

6.6.1 An unaltered copy of transferred data is retained by CRI and archived and managed with study data per the DMP, DUA, or contract requirements.

6.6.1.1 Received files are archives as received.

6.6.1.2 Data received through system interfaces are documented through system audit trail mechanisms.

6.6.2 CRI should not alter data for which an external party is responsible for data changes.

6.7 Quality checking transferred data

6.7.1 Data received by external parties shall be reconciled with study data to identify the following:

6.7.1.1 Received data not matching an existing study identifier

6.7.1.2 Study identifiers for which data were received that do not have a match within existing study identifiers

6.7.1.3 Study identifiers for which unexpected data were received

6.7.2 Additional consistency checking may be performed on externally managed data commensurate with the importance of the received data to the study and study scope of work.

6.7.3 Quality checking of transferred data may be documented in data transfer specifications (recommended so that external data providers know what to expect) or may be documented with other study data quality. Data quality rules will be subject to the version control applicable to their documentation.

7.0 SOP Deviations

Deviations from this and all SOPs are handled according to CRI.POL.001 *Clinical Research Informatics Quality Management System (QMS)*.

8.0 Review & Revisions

Review and revisions of this and all SOPs are handled according to CRI.POL.001 *Clinical Research Informatics Quality Management System (QMS)*.

9.0 Attachments

Attachment 1 Data Integration Form

10.0 **Revision History** (Since Last Version)

The revision history will be documented using the table shown below:

Section	Revision Date	Description of Revision

Attachment 1: Data Integration Form or Template (CRI.SOP.DMLC-003.FRM-001)

This form is used to document all externally originated or managed data for a study other than data key entered into eCRFs.

Section 1: Action initiated with this form (check one):

□ Initial version of this form Date: ___ ___ / ___ ___ ___ / ___ ___ ___
dd mon yyyy

□ Amendment to the initial version Date: ___ ___ / ___ ___ ___ / ___ ___ ___
dd mon yyyy

If applicable, reason for amendment:

Section 2: Sources of external data for the study:

List all sources of data for the study, whether or not managed by CRI. This list shall be comprehensive and match data sources indicated as external in section 2 of the Study Data Collection and Processing Plan Form (CRI.SOP.DMLC-004.FRM-002). Add additional instances of any choice below. Section 3 should be repeated and completed for data source indicated here.

□ Core lab: _______________________ or □ N/A

□ FHIR® based EHR-to-eCRF: _______________________ or □ N/A

□ Central lab: _______________________ or □ N/A

□ Central reading center: _______________________ or □ N/A

□ Electronic Clinical Outcomes Assessment (eCOA) Vendor: _____________________ or □ N/A

□ Claims data provider: _______________________ or □ N/A

□ Follow-up call center: _______________________ or □ N/A

□ Data transfer from study sites: _______________________ or □ N/A

□ Other: _______________________ or □ N/A

Section 3: Handling of External Data

The following items shall appear for each externally originated or managed data source.

A. **External Data Source** (*from section 2*): ___________________________________

B. **Blinding**:

□ The data are **not** blinded.

□ The data are blinded. *Unblinded individuals are listed below and should include members of the study team as well as roles of blinded site personnel, patients and their caregivers*.

C. **Informed Consent and HIPAA Authorization**:

□ I have confirmed that the external data supplier is listed in the informed consent or HIPAA Authorization or that informing research participants or their LAR is not required.

D. **Timing of data transfer and integration**:

Data will be sent by the external data provider and received by CRI on the following schedule. Describe the frequency if applicable and any specifics of the agreed schedule.

□ I have confirmed that the frequency and modality is consistent with CRIs and the external data provider's scope of work where applicable.

E. Data transmission mechanism

State the data transfer mechanism such as CRI sFTP, other UTHSA sFTP, external data providers sFTP. encrypted email, or a system interface.

□ I have confirmed that the data transfer mechanism is consistent with CRIs and the external data provider's scope of work where applicable.

F. Data storage

Data will be stored in the following locations upon receipt and prior to integration:

Received data will be stored in the following location: __________________________

Integrated data will be stored in the following location: __________________________

G. Data Exchange

The following exchange standard will be used: _______________________________

The following content standard will be used: _______________________________

□ Some or all transferred data or aspects of the transfer will not be covered by data exchange or content standards and data transfer specifications are required. If this box is checked, data transfer specifications are expected in the Data Management Plan. An external data provider's transfer specification format may be used. A plain text example file with all columns or tags defined and data types specified for each with valid values stated for discrete fields and decimal location and dimensionality specified for continuous data elements.

H. Data matching and linkage

The following data elements will be used to match incoming data values to other study data.
The data elements listed should be part of the study data dictionary and included in data transmitted by the external data provider. Quality checking of match data elements is strongly suggested. The listed data elements must provide for one-to-one matching of transmitted data to study data across the time-span of the study.

I. Exception checking and handling

Indicate the party/ies responsible for checking the consistency or other quality aspects of the incoming data. *Add rows under each heading to specify different types of checks. Add check boxes as appropriate.*

	EDP	CRI	Biost.	Oth.	**Tracks**	**Makes Updates***
Quality checking prior to data transfer	□					
Quality checking prior to import		□				
All files present		□				
File format and data type checks		□				
Patient matching (orphan records)		□				
Quality checking after import		□				
Identifier consistency checks		□				
Clinical data consistency checks		□	□			

EDP: External Data Provider.
*The party making updates to the data is responsible for maintaining the audit trail of such changes.

Statement of how exceptions will be communicated to the resolving party.

Statement of timeline within which resolutions are expected.

□ I have confirmed that the exception checking and handling tasks and associated audit trail maintenance responsibility are consistent with the scope of work for the indicated party/ies.

□ The external data provider's scope of work requires 21 CFR Part 11 compliance of them.

J. Acceptance criteria for external data

The criteria for final acceptance of data from the external data provider are listed below.
The criteria should be cross-listed on or referenced by the Database Lock checklist.

□ I have confirmed that these acceptance criteria that must be met prior to database lock are consistent with the external data provider's scope of work.

Repeat A-J for each external data source indicated in Section 2.

Approvals:

CRIS: ______________________________________

Signature: ____ Date: ___ ___ / ___ ___ ___ / ___ ___
dd mon yy

Study Statistician: ______________________________________

Signature: ____ Date: ___ ___ / ___ ___ ___ / ___ ___
dd mon yy

Study PI: ______________________________________

Signature: ____ Date: ___ ___ / ___ ___ ___ / ___ ___
dd mon yy

Signatures indicate review and agreement that all study data sources are listed on the form and approval of stated plans for CRIs handling of externally originated or managed data.