Proposed data standards for clinical and preclinical data produced by RMIP Investigators

BDC-Draft-#: 24
Title: Data Standards for Clinical and Preclinical Data
Type: Consensus Building
Contact Name and Email: Jeran Stratford, jstratford@rti.org
Submitting Team: NHLBI Data Management Core – Task 7 RMIP
Date Sent to Consortium: June 15, 2023
Status: Closed for comment
URL Link to this Google Document:
https://docs.google.com/document/d/1c8fIs5CITe6iHJjnQCkwbHDC14FuXfP83fidGEa3nGs/edit?usp=sharing

Goal of the document

The Regenerative Medicine Innovation Project (RMIP) spans research in diabetes, anemia, corneal and other ocular diseases, chronic skin ulcers, and rare diseases such as idiopathic pulmonary fibrosis, inherited skin diseases, and sickle cell disease. RMIP study data will be generated by RMIP investigators and the In-Depth Cell Characterization Hub, a component of the Regenerative Medicine Innovation Catalyst (RMIC). Intra- and inter-study interoperability of these data is essential to analyze the data efficiently and effectively. Employing data type specific standards provides a uniform framework that maximizes compatibility with existing tools and resources and the reuse potential of the data.

In this document we describe and request feedback for the proposed standards for clinical and pre-clinical data including phenotype data, clinical observations, and laboratory measurements. Standards regarding definition of raw data, file formats, terminology and vocabulary, required metadata associated with these data types will be discussed in greater detail below.

Table of Contents

1. Clinical Observations, Phenotypes, and Laboratory Measurements

1.1 Description and Use

Clinical trials are research studies that test a medical, surgical, or behavioral intervention in people and are the primary way that researchers determine if a new form of treatment or prevention is safe and effective in people. Various types of data are generated during a clinical trial including

- **Clinical observations** – refers to a result, answer, judgment, or knowledge gained from the act of observing a patient. This might be a vital sign measurement, a pain scale rating, or recording the kind of exercise activity a patient is engaged in. Depending on the study domain of interest, you might call these tests, variables, or data elements. Data sets may include baseline, interim visit, ancillary data, procedural based data, and outcome data.

- **Clinical phenotypes** - refers to the observable characteristics in an individual resulting from the expression of genes. Clinical studies collect measured or descriptive traits per individual participant. Linking clinical phenotypes to disease mechanisms can be used to define groups of patients who may or may not respond to a particular treatment.

- **Clinical laboratory measurements** – refer to measurements collected as part of the pre-clinical or clinical trial. These include data from assays and are reported per individual sample, where multiple samples may be associated with a single research participant. Results from test panels or abstraction of test results from medical records may be reported on a study form at the participant level, typically filled by study personnel or by the medical abstractor. Data may also include additional attributes about the sample important to understanding the study.

1.2 Data Collection

1.2.1 Clinical Data Standards

Use of data standards increases the quality of the data as well as making the data more interoperable allowing for combining study variables resulting in increased statistical power, enabling cross study comparisons to validate results, thus increasing the impact of individual studies. Therefore, data standards should be used whenever a relevant standard is available.

Many data standards are available for reporting clinical data and may be in the form of data models, common data elements (CDEs), ontologies and standardized vocabularies.

1.2.1.1 Data Models

A clinical data model is a framework for storing clinical information and defining the relationships between the collected data. A common data model provides an efficient means to integrate health data from multiple data sources as well as develop software tools that ensure reliable and reproducible research. Several commonly used data models include those supported by the following organizations: Informatics for Integrating Biology and the Bedside (i2b2), Observational Medical Outcomes Partnership (OMOP), Sentinel, and Patient Centered Outcomes Research Network (PCORnet) and the Study Data Tabulation Model (SDTM) developed by Clinical Data Interchange Standards Consortium (CDISC). **The CDISC SDTM model is the preferred model for reporting clinical trial data.**

1.2.1.2 Common Data Elements

Use of CDEs can help identify comparable datasets and reduce the need for retrospective harmonization. Investigators are encouraged to use CDEs for collection of clinical data. To search and explore relevant CDEs (e.g., LOINC, PhenX, caDSR), investigators can browse the NIH CDE repository or the caDSR CDE browser. CDEs should be specified within the data dictionary, in the VARIABLE_SOURCE and SOURCE_VARIABLE_ID columns (see below).

1.2.1.3 Ontologies and Standardized Vocabularies

Annotation of study variables with terms from controlled vocabularies is useful for cross-study comparisons. Ontologies, terminologies, and vocabularies have been created for consistent reporting of various clinical features including clinical diagnoses (e.g., ICD-9/ICD-10 codes, MONDO disease ontology), laboratory results (e.g., LOINC), medications (e.g., RxNorm, National Drug Code, DrugBank, DrugCentral), as well as general biomedical and health-related terms (e.g., MeSH, SNOMED, UMLS) To identify a relevant terminology or vocabulary, investigators can browse the Medical Subject Headings (MeSH), UMLS Metathesaurus, and the BioPortal search tool . **The recommended that the Mondo vocabulary be used to describe disease/phenotype terms, LOINC for laboratory results, RxNorm for medications, and Unified Code for Units of Measure (UCUM) for measurements. The recommended ontology for cell type terms is the Cell Ontology.** Controlled vocabulary terms should be specified within the data dictionary in the VARIABLE_SOURCE and SOURCE_VARIABLE_ID columns (see below).

1.2.2 RMIP Study Interoperability

To aid in interoperability or RMIP data we recommend collecting information about demographics, adverse events and outcomes using the following schemas:

1.2.2.1 Demographics

It is recommended that collection of participant demographic information will include the following data elements

- First, Middle, and Last Name
- Name Suffix
- Date of Birth
- Race
- Ethnicity
- Tribal Affiliation
- Biological Sex
- Sexual Orientation
- Gender Identity
- Preferred Language
- Occupation
- Contact Information including current address, phone number, and email address

Collected data should be formatted to comply with the CDISC demographics standard. A description of data elements to capture this information has been provided in Appendix A. Additional demographic information may be added to this minimum set as appropriate to meet the needs of the study but will need to be included and described in the data dictionary.

While we recommend that this demographic information is collected, certain fields will need to be omitted in files uploaded to BioData Catalyst. See section below on de-identification.

1.2.2.2 Adverse Events

It is recommended that collection of adverse event information will include the following data elements:

- Adverse event start and end date

- Uncoded description of the adverse event
- Coded term describing the adverse event (e.g., MedDRA)
- Standard toxicity grade (e.g., CTCAE scale)
- Whether the adverse event is considered a serious event
- Indicator whether the adverse event is ongoing
- Outcome of the adverse event including hospitalization or death
- Action taken
- Adverse event causality

Collected data should be formatted to comply with the CDISC adverse events standard. A description of data elements to capture this information is provided in Appendix B. Additional adverse event data elements may be added as appropriate to meet the needs of the study but will need to be included and described in the data dictionary.

1.2.2.3 Outcomes

A primary goal of RMIP is to enable the investigation of the correlation between clinical outcomes and the cell product attributes. To enable this, it is recommended that **investigators clearly define the criteria for a successful study/trial outcome** and provide a binary indicator for each participant describing whether the criteria for a successful study/trial outcome was satisfied. This definition will be supplied in the data dictionary file associated with this dataset.

The data set file will include elements for

- Participant ID
- Was the success criteria met for this participant?

A description of data elements to capture this information is provided in Appendix C. The preferred format for collected data is a delimited text file (.csv or .tsv), but an MS Excel Workbook (.xlsx) is also acceptable.

1.3 Reporting Clinical Data

The following files are required for submission of clinical data for RMIP investigations:

1. Dataset (DS) File
2. Data Dictionary (DD) File
3. Sample Manifest

Clinical information is reported in a dataset (DS) file. A DS file is a rectangular table of data values, subject/sample IDs and variables. A single DS file is typically provided for each domain/form for a study (e.g., demographics, medications, outcomes) and typically contain participant-level data for each study variable. DS files are discussed in greater detail below.

DS files are always accompanied by a data dictionary (DD) file. A data dictionary is used to catalog and describe the structure and content of collected data reported in a DS file and are useful by defining project conventions and establishing data collection consistency across multiple research team members and sites. DD files are discussed in greater detail below.

1.3.1 Dataset File

The Dataset file includes measured and/or descriptive traits per individual person or per sample for clinical laboratory data. The primary ID in this file is the SUBJECT_ID. Each dataset (DS) file must be submitted with a corresponding DD file. Example templates for phenotype and laboratory data. We recommended reporting the data in a table with the following columns:

1) SUBJECT_ID or PARTICIPANT_ID - Each SUBJECT_ID needs to be unique and should be linked to only 1 row of data in the DS. All SUBJECT_IDs included in this file must be from consented subjects.

2) All other Column Headers: VARNAMES (variable names) – Submit data relevant to an individual person. Variable names should correspond with those included in the data dictionary. This may include variables like affection status or case/control status of the disease/phenotype, race/ethnicity/ancestry/heritage, and relevant dates written as years or normalized to a set point in time. Do not include month and days directly tied to the person, which are considered HIPAA sensitive.

For laboratory measures, submit sample attribute variables that will provide a greater understanding of the study. Additional variables can be added if needed, for example, collection site of the sample, analyte type of the sample, or the sample's cell or tissue type/subtype. Relevant dates (e.g., sample collection date) that are directly tied to a person should be written as years or normalized to a set point in time.

1.3.1.1. Format

A separate dataset file is expected for each data domain (e.g., demographics, laboratory results, health care utilization) and there is no limit on the number of domains that can be submitted but is dictated by the study design. Dataset files will contain tables with one row for each subject or one row for each individual sample. The table should be submitted either in delimited text file (.csv or .tsv) or a Microsoft Excel workbook (.xlsx), with a delimited text file being the preferred format.

1.3.2 Data Dictionaries

In the data dictionary a meaningful description of each individually named variable is provided and typically includes information about:

1. The name of the collected variable (e.g., AGE, NEUTROPHILS,)
2. A description of what information was collected
3. Questions used or description defining how the data was collected
4. The data type (e.g., character, numeric, integer, float, Boolean, …)
5. Format of data, which may include specific syntax (e.g., date format YYYY-MM-DD), available responses (e.g., Yes|No, Poor|Good|Excellenet), data encoding (e.g., 1 = Female | 2 = Male), and the size (e.g., character length, number of decimal places reported)
6. Relationship of variables (e.g., defining how data may contribute to other derived or summary variables)

1.3.2.1 Format

The data dictionary should be submitted in either .txt or .xlsx format, with .xlsx being the preferred format. As a data dictionary defines and describes the variables in the corresponding dataset file, there will typically be one data dictionary for each dataset file provided. It is preferred that a single MS Excel workbooks (.xlsx) be submitted with each dataset dictionary provided on a separate worksheet.

There are many features that could be reported for each variable. A complete list of appropriate data dictionary descriptions and specifications is available for review from dbGaP with an example template.

The recommended columns for inclusion in a data dictionary are listed below and required fields are marked with an asterisk.

Column Headers	Description
VARNAME*	Variable name. The VARNAME must not contain backward slashes (\).
VARDESC*	Variable description. The description should be understandable and enable users to replicate the variable. For example, "blood pressure" is useful, but "brachial blood pressure while sitting" provides more context. Alternatively, study documents with detail may be referenced.
DOCFILE*	Study document name associated with the variable. To list multiple documents, add a semicolon (;) between documents.
TYPE*	Data value type: integer (1,2,3,4,…), encoded value (integers or strings are coded for non-numerical meaning, ex. 1=Control; 2=Case, see VALUES), decimal (0.5,2.5,…), string (African American, Asian, Caucasian, Hispanic, Non-Hispanic). For mixed values (any combination of string, integers, decimals and/or encoded values) in a single data column, list all types present.
UNITS*	Units of measurement of variable
MIN	The logical minimum value of the variable. If a separate code such as -1 is used for a missing field, this should not be considered as the MIN value.
MAX	The logical maximum value for the variable. If a separate code such as 9999 is used for a missing field, this should not be considered as the MAX value.
RESOLUTION	Measurement resolution – the number of decimal places to which a measured value is presented in the data. For example, in 54.321 the resolution is 3.
COMMENT1, COMMENT2	Additional information not included in the VARDESC that will further define the variable. If additional comments are needed beyond COMMENT2, insert new columns (COMMENT3, COMMENT4, etc.) before the column "ORDER."
VARIABLE_SOURCE	Source of controlled vocabularies. Ex. PhenX, MeSH, SNOMED, NCI. If there is no match, leave blank. (Must be submitted as a group with SOURCE_VARIABLE_ID and VARIABLE_MAPPING).
SOURCE_VARIABLE_ID	A unique identifier from the VARIABLE_SOURCE or a unique text concept/term from various controlled vocabularies. (Must be submitted as a group with VARIABLE_SOURCE and VARIABLE_MAPPING).

VARIABLE_MAPPING	For example, a variable from the source could be Identical, Related, or Comparable. (Must be submitted as a group with VARIABLE_SOURCE and SOURCE_VARIABLE_ID).
COLLINTERVAL	Collection interval is the time frame in which the data for the variable or dataset was collected.
VALUES*	List of all unique values and/or descriptions of all encoded values, one value per cell. Encoded values are defined as a value and its meaning. For example, if a data file contains a variable named "EDUCATION" and its data values are "1, 2, 3, and 99," these coded values will need to be defined in the data dictionary. The format of an encoded value is VALUE=MEANING. Therefore, in the data dictionary, there should be 4 separate data cells filled out with the following: 1=Completed High School, 2=Completed College, 3=Completed Graduate School, 99=Unknown. The "VALUES" header must be the last column header (farthest right in the table). It should appear only in the column above the first encoded value that is listed. The remaining column header cells should be left blank. The script will identify the first code meanings and continue right until there are no more code meanings. For example, if the variable "SEX" has 3 encoded values: 1=Male, 2=Female and 3=Unknown, the column header "VALUES" will appear only above the cell that contains 1=Male. 1=Male, 2=Female and 3=Unknown will be listed in three separate cells next to each other. The header column cells above "2=Female" and "3=Unknown" should be left blank.

1.3.3 Sample Manifest

The sample manifest is used to catalog and describe the collection of biospecimens (origin, collection mechanism, timing, quantity) and link them to specific study participants. Each study should have a single sample manifest that contains information about the collection of all biospecimens. This document provides the essential information for linking together research participant outcomes with the files that contain cell characterization profiles.

1.3.3.1 Format

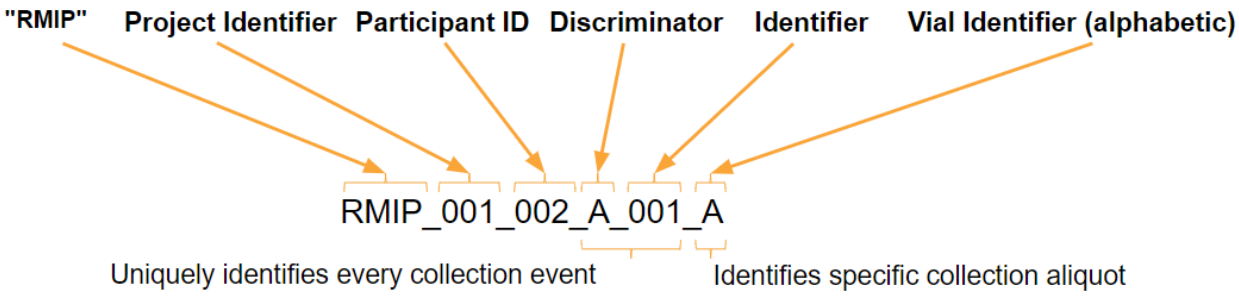
The sample manifest should be submitted in either .txt or .xlsx format, with .xlsx being the preferred format.

The recommended columns to be included in the sample manifest are listed below and required fields are marked with an asterisk.

Column Headers	Description
PROJECT*	Project identifier. The three-digit ID assigned to each RMIP study by the in-depth cell characterization hub.
PARTICIPANT*	Participant ID. A de-identified accession number describing a single research participant assigned by the investigator.
DISCRIMINATOR*	Collection event identifier part 1. Alphabetic component of the sample discriminator combination that uniquely identifies each sample collection event.

IDENTIFIER*	Collection event identifier part 2. Numeric component of the sample discriminator combination that unique identifies each sample collection event.
TIMEPOINT	Collection time. Description of when the sample was collected, relative timepoints are encouraged (e.g., baseline, 6 months follow-up, study end) but collection dates are also acceptable but may require de-identification (date shifting) to protect participant privacy.
SOURCE	Biospecimen source. Description of the material that collected.
METHOD	Collection Method. Method used to collect the sample.
PROCESSING	Sample processing. Any methods used to process or refine the collected sample.
REPLICATES*	Collection replicates. The number of replicate collections or aliquots that were obtained during the event. Replicate biospecimens will each be assigned an alphabetic value as a vial identifier.

Information contained in the sample manifest is used to generate an identifier that will be used as a filename prefix for all data that is generated from the specific biospecimen collection. The collection ID has the following structure



1.3.4 Important Considerations

1.3.4.1 De-identification

To comply with HIPAA, personally identifying information must be removed from all data, e.g., names, cities, dates, telephone numbers, social security numbers, and any other potentially identifying information, characteristic, or code.

The de-identified participant ID (either random or consecutive numbers) should be assigned to all participants by the investigator. This participant ID is the primary key that should be used throughout all dataset files when reporting data. A mapping file defining which ID corresponds to each participant should be created and maintained by the investigator but not shared or uploaded to BioData Catalyst.

1.3.4.2 Reporting Dates

There are many different ways to report dates. It is recommended that all dates be reported in the ISO 8601 date format of YYYY-MM-DD.

2. Preclinical Data

2.1 Description and Use

Before a new form of treatment or prevention can be tested in people, researchers must determine whether there are potential safety concerns and whether the approach is likely to be useful. Preclinical studies often involve the use of animal models to evaluate the *in vivo* efficacy and safety of the approach. The choice of animal model and data to be collected are dependent on the study being conducted. Data generated from preclinical studies often include content similar to what has been previously described for clinical trials.

2.2 Data Collection

2.2.1 Preclinical Data Standards

Many of the standards discussed for clinical data are also applicable for preclinical studies. It is recommended that relevant data models, vocabularies and ontologies be used when available. When possible, please use standards that apply broadly to both human and animal studies. Identification of available CDEs, ontology, and vocabulary standards for preclinical studies can be accomplished using the search tools previously indicated for clinical data.

2.2.2 Demographics

When animal models are involved with preclinical research it is essential to report the characteristics of all animals. This is equivalent to the standardized demographic data for human research participants, enabling researchers to repeat the experiment, generalize the findings, and assess whether the animal characteristics are relevant to the research objectives. **It is recommended that the species, strain, substrain, sex, weight, and age of the animal be recorded.** As body weight may vary over the course of the study, it is important to record the weight whenever measurements are taken and indicate when the data was collected. When possible, baseline age and weight values for individual animals should be provided.

2.4 Reporting Preclinical Data

The same dataset (DS) and data dictionary (DD) reporting structure outlined for reporting clinical data should be utilized for reporting preclinical data.

Appendix A:

Fields for collecting demographics data

Variable Label	Variable Name*	Variable Description	Values
Subject ID	SUBJID	Subject Identifier for the Study	[SUBJECT ID]
Demographics Collection Date	DMDAT	Date that the demographic information was collected	[DATE: YYYY-MM-DD]
First Name	--	Subject first name	[Free Text]
Last Name	--	Subject last name	[Free Text]
Middle Name	--	Subject middle name (including middle initial)	[Free Text]
Name Suffix	--	Relevant language that follows someone's name used to identify the person (e.g., Jr, II)	[Free Text]
Date of Birth	BRTHDAT	What is the subject's date of birth?	[DATE: YYYY-MM-DD]
Race	RACE	Which of the following five racial designations best describes you? (More than one choice is acceptable.)	American Indian or Alaska Native | Asian | Black or African American | Native Hawaiian or Other Pacific Islander | White | Not Reported | Unknown | Other
Specify Other Race	RACEOTH	What was the other race?	[Free Text]
Ethnicity	ETHNIC	Do you consider yourself Hispanic/Latino or not Hispanic/Latino?	Hispanic or Latino | Not Hispanic or Latino | Not Reported | Unknown
Tribal Affiliation	TRIBE	What is the name of the Tribal Nation that the subject is affiliated with?	[Free Text]
Sex	SEX	What is the sex of the subject?	Female | Male | Unknown | Undifferentiated
Sexual Orientation	SEXIDENT	Which of the following best represents how the participant thinks of themselves?	Gay | Lesbian | Straight; That is, not gay or lesbian, etc. | Bisexual | Unknown | Prefer not to answer
Gender Identity	GENIDENT	What terms best express how the participant describes their gender identity?	Man | Woman | Non-Binary| Transgender | Unknown | Prefer not to answer
Preferred Language	LANGPREF	What is the participant's preferred language to speak and write?	English | Spanish | Chinese (Mandarin or Cantonese) | Tagalog | Vietnamese | Arabic | French | Other
Street Address	--	What is the participant's full street address for their residence?	[Free Text]

Phone Number	--	What is the participant's telephone number where they can be reached?	[Free Text]
Phone Number Type	--	What type of phone is	Mobile | Home | Work | Other
Email Address	--	What is the participant's email address?	[Free Text]
Occupation	JOB	What is the participants occupation?	[Free Text]
Occupation Industry	JOBCLAS	What kind of business or industry is this? (For example: a TV or radio station, retail shoe store, state labor department, farm.)**	[Free Text] ENTER NAME OF BUSINESS OR INDUSTRY
Education Level	EDULEVEL	What is the highest grade or level of school you have completed or the highest degree you have received?**	NEVER ATTENDED/KINDERGARTEN ONLY | 1ST GRADE | 2ND GRADE | 3RD GRADE | 4TH GRADE | 5TH GRADE | 6TH GRADE | 7TH GRADE | 8TH GRADE | 9TH GRADE | 10TH GRADE | 11TH GRADE | 12TH GRADE, NO DIPLOMA | HIGH SCHOOL GRADUATE | GED OR EQUIVALENT | SOME COLLEGE, NO DEGREE | ASSOCIATE DEGREE: OCCUPATIONAL, TECHNICAL, OR VOCATIONAL PROGRAM | ASSOCIATE DEGREE: ACADEMIC PROGRAM | BACHELOR'S DEGREE (EXAMPLE: BA, AB, BS, BBA) | MASTER'S DEGREE (EXAMPLE: MA, MS, MEng, MEd, MBA) | PROFESSIONAL SCHOOL DEGREE (EXAMPLE: MD, DDS, DVM, JD) | DOCTORAL DEGREE (EXAMPLE: PhD, EdD) | REFUSED | DON'T KNOW
Marital Status	MARISTAT	What best describes your current marital status?**	Married | Divorced | Widowed | Separated | Never married | A member of an unmarried couple | Refused

* A variable name reported as "--" indicates that this variable should be collected by the study but that this variable should not be included in the data uploaded to BioData Catalyst.

**Common data element provided in the PhenX toolkit:

- PhenX Occupation/Occupational History, PX060501040000

- PhenX Educational Attainment – Individual, PX011002010000
- PhenX Current Marital Status, PX010903000000

Appendix B:

Fields for reporting adverse event data

Variable Label	**Variable Name**	**Variable Description**	**Values**
Subject Identifier for the Study	SUBJID		[SUBJECT ID]
Adverse Event Start Date	AESTDAT	What is the adverse event start date?	[DATE: YYYY-MM-DD]
Adverse Event End Date	AEENDAT	What is the adverse event end date?	[DATE: YYYY-MM-DD]
Adverse Event Term	AETERM	What is the adverse event term (uncoded)?	[FREE TEXT]
Adverse Event Code	AEDECOD	Provide the MedDRA term describing the adverse event.	[MEDDRA TERM LIST]
AE Standard Toxicity Grade	AETOXGR	What was the NCI CTCAE grade of the adverse event?	GRADE 1 | GRADE 2 | GRADE 3 | GRADE 4 | GRADE 5
AE Serious Event	AESER	Was the adverse event serious?	YES | NO
Ongoing Adverse Event	AEONGO	Is the adverse event ongoing at end of study?	YES | NO
Results in Death	AESDTH	Did the adverse event result in death?	YES | NO
Requires or Prolongs Hospitalization	AESHOSP	Did the adverse event result in initial or prolonged hospitalization of the participant?	YES | NO
Life Threatening	AESLIFE	Was the adverse event life threatening?	YES | NO
Action Taken with Study Treatment	AEACN	What action was taken with study treatment?	DOSE INCREASED | DOSE NOT CHANGED | DOSE REDUCED | DOSE RATE REDUCED | DRUG INTERRUPTED | DRUG WITHDRAWN | NOT APPLICABLE | UNKNOWN
Outcome of Adverse Event	AEOUT	What is the outcome of the adverse event?	FATAL | NOT RECOVERED/NOT RESOLVED | RECOVERED/RESOLVED | RECOVERED/RESOLVED WITH SEQUELAE | RECOVERING/RESOLVING | UNKNOWN
AE Causality	AEREL	Was this adverse event related to study treatment?	UNRELATED | POSSIBLY RELATED | RELATED

Appendix C:

Fields for reporting outcome data

Variable Label	**Variable Name**	**Variable Description**	**Values**
Subject Identifier for the Study	SUBJID*		[SUBJECT ID]
Original Result	TEST	Description of the success criteria	[Free text]
Text description	TESTRES	Description of the outcome for the participant?	[Free text]
Success criteria met	AVAL*	Was the defined success criteria met?	True | False