Paper DH05

Demystify the Handling and Mapping of External Data to CDISC Standards in Clinical Trials

Vamsi Kandimalla, Kite Pharma Inc, a Gilead Company, Santa Monica, USA
Murali Kanakenahalli, Kite Pharma Inc, a Gilead Company, Santa Monica, USA

Abstract

As the complexity of clinical trials has grown over the years, the dependency on different kinds of external data has increased tremendously. Different external data types like pharmacokinetic, immunogenicity, biomarker, genomics, cell phenotype, biospecimen, microbiology findings, etc., have been used to support the primary, secondary end points of a study and exploratory analysis. It's a substantial task for statistical programming teams to understand, analyze and maintain external data. A great deal of resources is required to develop an infrastructure to process the external data along with using CDISC implementation to map to SDTM standards.

We will discuss the various types of external data. Use data transfer plan with labs to receive clean data transfers. Understand external data and use CDISC to map to SDTM. Reconcile external data and handle some of the data reconciliation challenges. Also, discuss the documentation which can be reference in understanding and mapping to SDTM.

Introduction

The rapidly advancing new drug development across various therapeutic areas prompt to perform variety of clinical data analysis and reporting. Of which the external data from a lab by analyzing the collected subject's samples is an extensive area of interest. In other words the non-CRF data.

In this paper we have outlined an approach on how to handle and map the external data which we exercised and benefited.

- Sources: Explore and gather the essential information from sources available to understand these data.
- DTP: Design on how to setup a data transfer plan (DTP) with a lab to include the required and expected data points in support of downstream analysis reporting and as well to remain compliance with implementation guide (IG). Designing a DTP is a cross functional effort and responsibility involving data management, programming, biostatistician, medical and clinical teams.
- Map: The information collected from available sources, and DTP and then identifying the data points like type of sample collected, objective of sample analysis, type of sample analysis, type of method and modality, type of results generated and its downstream use, etc will assist to understand the scope of a particular external data to process it and map it to CDISC/SDTM standards.
- Quality Control: Data transfers per DTP, reconciliation, data issue identification and resolution, and maintain CDISC compliance.
- Track: Maintain a tracking tool as a documentation reference for consistency and standardization.

There are various types/examples of external data that gets processed, mapped and reported alongside CRF collected data. The common data are pharmacokinetic, immunogenicity, biomarker, genomics, cell phenotype, microbiology findings, etc. To detail the above mentioned approach for these listed common external data, we have addressed it by giving a brief introduction about each of the data examples, how the data is reported downstream, how to design a DTP, and how to map and track it to maintain consistency.

Based on SDTM IG 3.4, some of the individual finding domains that are designed to submit the external lab/stake holder analyzed data are PC, IS, MI, GF (formerly PF), CP, BS, MB, etc; or a possibility for a custom domain as appropriate.

This paper can help clinical and statistical programmers in ways to more efficiently explore these data, understand and structure it for analysis and reporting.

Approach to Handle and Map External Data

Sources:

- Protocol, statistical analysis plan (SAP), lab provided charters and documentation, lab's web portal, translational and statistical teams (scientist, statistician who requests this type of data for analysis needs), medical and clinical teams, understanding the analysis requirement (on how study design plan to present the data – tables, figures, listings), understanding the submission requirement (clinical study report (CSR) or exploratory purposes or publication), and other authenticate online resources, etc.

Examples of External data in Clinical Trials

Pharmacokinetic Data:

Pharmacokinetics (PK) is the study of how body interacts with administered substance for the duration of the exposure. Pharmacokinetics attempts to summarize the movement of drugs throughout the body and the actions of the body on drug. By using theories and equations, practitioners can better estimate the location and concentration of a drug in different areas of the body. The concentration needed to obtain the desired effect and the amount required for a higher chance of adverse reactions are determined through laboratory testing. Some of the tests are Polymerase Chain reaction (PCR) and Flow cytometry.

PCR and Flow cytometry data is an external data which falls under PK data. PCR is a method to amplify specific copy of DNA to millions or billions of copies, allowing scientist to use small copy of DNA for their detailed research. Flow cytometry is a technique used to detect and measure physical and chemical characteristics of population of cells. PCR and Flow cytometry is routinely used in cancer research. Flow cytometry is used for cell counting, biomarker detection, diagnosis of blood cancers etc. Figure 1 explains how the analyte concentrations are summarized in the table.

Figure 1: Generic summary shell for PCR/ Flow cytometry test and its analyte.

Summary of XXX (Cells/ug DNA) in ddPCR
Safety Analysis Set

	Treatment 1 (N = X)
Baseline	
n	x
Median (Q1, Q3)	x.x (x.xx, x.xx)
Min, Max	x.xx, x.xx
Post Baseline time point 1	
n	x
Median (Q1, Q3)	x.x (x.xx, x.xx)
Min, Max	x.xx, x.xx
Area Under Curve (from Day x to Day x+15)	
n	x
Median (Q1, Q3)	x.x (x.xx, x.xx)
Min, Max	x.xx, x.xx
Time to Peak (in Days)	
n	x
Median (Q1, Q3)	x.x (x.xx, x.xx)
Min, Max	x.xx, x.xx

Figure 2: Generic listing shell for PCR/ flow cytometry test and its analyte.

XXX (Cells/ug DNA)

Safety Analysis Set

	Treatment 1	Treatment 2	Treatment 3
Baseline	x.xx	x.xx	x.xx
Postbaseline Timepoint 1	x.xx	x.xx	x.xx
Postbaseline Timepoint 1	x.xx	x.xx	x.xx
Peak Concentration	x.xx	x.xx	x.xx
AUC (Day x to Day x+15)	x.xx	x.xx	x.xx
TTPEAK	x.xx	x.xx	x.xx

Once we understand the data details and how it is presented, next step is to design on how this data will be collected. Data transfer plan (DTP) helps as a living document which helps to collect receive data about PCR or flow cytometry. Programming teams work with Data management teams to create and review DTP to build an expectation of external data receival/collection. Generic information expected for PCR external data collection is shown below.

Table 1: Example of a DTP for PCR data

Field Name	SAS Name	Length	Data Type	Explanation	Example
File Creation Date and Time	FILCRDTM	50	Text	The date/time file created by sender	DDMONYYYY XX:XX
External Vendor Name	LABNAM	50	Text	*Vendor name XXX*	
Protocol	STUDNAM	25	Text	Protocol number	
Site ID	SITEID	3	Text	Site number	001, 002, etc.
Subject ID	SUBJID	11	Text	Subject number	001-001-001 etc.
Visit Name	VISIT	40	Text	Visit name explained in DTP	
Collection Date	LBDT	11	Text	Collection date	DDMONYYYY
Visit Type	VISITTYP	1	Text	Planned visit and unscheduled	Day 1, Unscheduled
Battery ID	BATTID	20	Text	The ID of the Battery group to which the panel belongs	PCR
Battery Name	BATTNAM	20	Text	The name of the Battery group to which the panel belongs	xxx cell DNA
Reported Result	RPTRESC	40	Text	Reported Result	
Reported Result Unit	RPTU	20	Text	Reported Unit	
Sample Type	SPEC	20	Text	Sample analyzed	PLASMA
Test Status	TSTSTAT	1	Text	D = Done, N = Not Performed, X = Cancelled	D, N, X
Comment	CMNT	200	Text		

Map: PCR and Flow cytometry tests provide the concentrations of drugs or metabolites. This is the primary reason PCR/Flow cytometry data is mapped to PC domain. The assumptions which help us to decide to map PCR/flow cytometry data into PC domain are:

- PC domain can be used to represent specimen properties like quantity count, volume in addition to drug and metabolite concentration.
- Identifiers variable, time variables and findings class are usually involved in PCR/flow cytometry data.

Considering the raw data type and assumptions an example of how the raw dataset values can be mapped to PC domain is shown below.

- PCCAT – can contain values like “Quantitative Assay”.
- PCTEST – the actual PCR test.
- PCORRES/PCSTRESN/PCSTRESC – the result/finding of the PCR test.

- PCORRESU/PCSTRESU – unit of the result/finding. Example: if 1000 is the result, units can be in copies per microgram of DNA.
- PCSPEC – type of specimen like “PLASMA”.
- VISIT/PCDTC – Timing variables of the PCR test performed.

Immunogenicity Data:

Immunogenicity is the ability of the drug or the antigen to cause immune response in the body. The immune system recognizes antigens on the surface of cancer cells as foreign substance. The assessment that determines whether a drug induced an immune response falls under findings domain. Some of the tests like Antibody screening or confirmation, B cell aplasia fall under immunogenicity domain.

Immunogenicity evaluation is typically evaluated using antibody screening and confirmatory tests in a tiered manner. Antibody screening test is used to detect all antibodies bind to the drug. The Antibody confirmatory tests are conducted to confirm the antibody screening results. FDA also has guidance on Immunogenicity testing of therapeutic protein products for Anti-drug antibody detection. Using flow cytometry, a blood test is performed to measure the number of B lymphocytes in the blood. With significantly low count indicates B cell aphasia. In a clinical context this test is particularly important for patients undergoing CAR-T cell therapy.

Data collection can be understood using the DTP for antibody screening/confirmation test and B cell aphasia test.

Table 2: Example of a DTP for antibody screening/confirmation data.

Field Name	SAS Name	Length	Data Type	Explanation	Example
File Creation Date and Time	FILCRDTM	50	Text	The date/time file created by sender	DDMONYYYY XX:XX
External Vendor Name	LABNAM	50	Text	*Vendor name XXX*	
Protocol	STUDNAM	25	Text	Protocol number	
Site ID	SITEID	3	Text	Site number	001, 002, etc.
Subject ID	SUBJID	11	Text	Subject number	001-001-001 etc.
Visit Name	VISIT	40	Text	Visit name explained in DTP	
Date of Specimen Collection	COLLDAT	15	Text	Date actual sample was taken from subject	DDMONYYYY
Accession Number	ACCSNNUM	15	Text	Central Lab Accession Number (Sample ID)	ABC0123456, etc.
Results	RSLT	50	Text	Positive =>1.58 Norm MFI (median fluorescent intensity) Negative <1.58 Norm MFI	Positive Negative No Result
Sample Norm	SAMPNORM	20	Text	Sample normalized MFI	3.4, 1.3, etc. May Be Blank
Sample Norm Unit	NORMU	20	Text	Normalized median fluorescent intensity	Norm MFI May Be Blank

Table 3: Example of a DTP for B cell aphasia data.

Field Name	SAS Name	Length	Data Type	Explanation	Example
File Creation Date and Time	FILCRDTM	50	Text	The date/time file created by sender	DDMONYYYY XX:XX
External Vendor Name	LABNAM	50	Text	*Vendor name XXX*	
Protocol	STUDNAM	25	Text	Protocol number	
Site ID	SITEID	3	Text	Sit enumber	001, 002, etc.
Subject ID	SUBJID	11	Text	Subject number	001-001-001 etc.

Field Name	SAS Name	Length	Data Type	Explanation	Example
Visit Name	VISIT	40	Text	Visit name explained in DTP	
Sample Date	SAMPDAT	15	Text	Date sample was taken from subject	DDMONYYYY
Assay Date	ASSAYDAT	20	Text	Date sample was analyzed	DDMONYYYY
Assay Number	ASSAYNO	20	Text	Unique URMC sample ID number	000654321, etc.
Analyte Name	ANALYTE	50	Text	Analyte	% B Cells of Viable Leukocytes
Results	RSLT	50	Text	Numerical, BLQ or ND	0.02, BLQ, ND
Units	UNIT	40	Text	Unit for Numerical Result	%

Map: The assumptions which help to put the test on antibody screening/conformation, B cell aphasia are:

- Since these tests are about the assessments what describe whether a therapy induced an immune response. IS domain model holds these assessments.
- Identifier variables, timing variables and findings variables are collected in these assessments.

Considering the assumptions, an example of how the raw data information is mapped to IS domain variables is shown.

- ISCAT – contains values like 'ANTIBODY CONFIRMATORY RESULT' 'PERIPHERAL BLOOD MONONUCLEAR CELL'.
- ISTEST – the actual test like 'Antibody Test Confirmatory', '% B Cells of Viable Leukocytes'.
- ISORRES/ISSTRESN/ISSTRESC – the result/finding of the immunogenicity test.
- ISORRESU/ISSTRESU – unit of the result/finding.
- ISNAM – name of the vendor where the test was performed.
- VISIT/ISDTC – Timing variables of the immunogenicity test performed.

Biomarker Data:

Biomarker data in clinical trials refers to measurable biological characters like proteins or genetic markers collected from subjects during a clinical trial to evaluate how they are responding to treatment, and allowing for any early detection of positive or negative effects.

Table 4: Example of a DTP for a Microscopic data - Immunohistochemistry (IHC) data

Field Name	SAS Name	Length	Data Type	Explanation	Example
File Creation Date and Time	FILCRDTM	50	Text	The date/time file created by sender	DDMONYYYY XX:XX
External Vendor Name	LABNAM	50	Text	*Vendor name XXX*	
Protocol	STUDNAM	25	Text	Protocol number	
Site ID	SITEID	3	Text	Site number	001, 002, etc.
Subject ID	SUBJID	11	Text	Subject number	001-001-001 etc.
Visit Name	VISIT	40	Text	Visit name explained in DTP	
Collection Date	LBDT	11	Text	Collection date	DDMONYYYY
Visit Type	VISITTYP	1	Text	Planned visit and unscheduled	Day 1, Unscheduled
Accession Number	ACCESNUM	10	Text	Accession number	1234567
Sample Type	SPEC	20	Text	Sample type	Bone Marrow, Other, Unknown
IHC Test	TEST	40	Text	IHC tests	CD6, CD7, etc
Percent Positive	PPOS	4	Text	Percentage of stained cells	0-100, NA
Percent Positive Units	PPOSU	2	Text	Percentage of stained cells unit	%
Staining Intensity	STINT	10	Text	Staining intensity of cells	0+, 1+, 2+, 3+
H-Score	HSCORE	10	Text	Cumulative H-score for cells	0-300
FISH Result	FISHRSLT	40	Text	FISH result	Normal,

Field Name	SAS Name	Length	Data Type	Explanation	Example
					Abnormal etc
Histological Confirmation	HISTCONF	100	Text	Type of Disease	TFL, MZL, Inconclusive
Comments	COM	200	Text	Comments	XXXX

In regards to the Table 4 listed DTP, the testing is performed on the samples to identify biomarkers for cancer diagnosis and treatment using IHC as a testing modality. This testing would help to form a complete picture for cancer diagnosis, make informed treatment decisions, provides insights on unknown subject population and for a personalized care. Examples of other testing modalities that are also used to perform the biomarker identification are fluorescence in situ hybridization (FISH), flow cytometry etc.

IHC is a method that involves treating tissue with a stain that adheres to specific substance. The Staining intensity of cells was assessed on a semi-quantitative scale ranging 0-3+ and percentage of cells at each staining intensity level were reported. These results were used to calculate the H-score that range from 0-300.

Map: The data here are findings based on microscopic examination. These biomarkers are assessed using histologic or histopathological examination done by using stains. This data aligns with the definition of SDTM IG's definition for Microscopic Findings domain (MI).

Considering the assumptions, below is the high level mapping of this example external raw data to MI domain.

- MITEST will have values 'CD5', CD7' etc
- MITSTDTL will capture values like 'HSCORE STAINING INTENSITY 0', 'HSCORE STAINING INTENSITY 1+', 'HSCORE STAINING INTENSITY 2+', 'HSCORE STAINING INTENSITY 3+', HSCORE TOTSL SCORE'
- MIORRES/MISTRESC/MISTRESN will capture the values from result field
- MIORRESU/MISTRESU will capture the values '%' from Unit field
- MIDTC will capture the value from collection date field
- MIMETHOD here are 'IHC', "FISH'
- MISPEC here is 'BONE MARROW'
- CO domain captures the Comments
- Any additional content not part of parent domain will be submitted under SUPP domain.

Genomics Data:

Genomic data in clinical trials refers to use of subject's genetic information obtained through DNA sequencing to understand disease mechanisms, predict treatment responses, design treatment based on subject's genetic profile.

Based on the SDTM IG 3.4, this type of genomics data is recommended to be submitted in a newly introduced genomics findings (GF) domain. This genomics data contain data related to structure, function, mapping, evolution and editing of subject and non-host organism genomic material of interest, and include results for genetic variation, transcription and summary measures. It is also includes characteristics assessed from nucleic acid and inferences and predictions about proteins/amino acids.

Table 5: Example of a DTP for Genomic Data – PanCancer IO 360™ panel data

Field Name	SAS Name	Length	Data Type	Explanation	Example
File Creation Date and Time	FILCRDTM	50	Text	The date/time file created by sender	DDMONYYYY XX:XX
External Vendor Name	LABNAM	50	Text	*Vendor name XXX*	
Protocol	STUDNAM	25	Text	Protocol number	
Site ID	SITEID	3	Text	Site number	001, 002, etc.
Subject ID	SUBJID	11	Text	Subject number	001-001-001 etc.
Visit Name	VISIT	40	Text	Visit name explained in DTP	

Field Name	SAS Name	Length	Data Type	Explanation	Example
Collection Date	LBDT	11	Text	Collection date	DDMONYYYY
Visit Type	VISITTYP	1	Text	Planned visit and unscheduled	Day 1, Unscheduled
Accession Number	ACCESNUM	10	Text	Accession number	1234567
ARG1	ARG1	10	Text	Gene expression score btw -5 to 10	-2.0000, 1.0000
CTLA4	CTLA4	10	Text	Gene expression score btw -5 to 10	-2.0000, 1.0000
IL10		10	Text	Gene expression score btw -5 to 10	-2.0000, 1.0000
Inflammatory Chemokines	INFL_CHM	10	Text	Gene expression score btw -5 to 10	-2.0000, 1.0000
Myeloid Inflammatory	MYLD_INF	10	Text	Gene expression score btw -5 to 10	-2.0000, 1.0000
NOS2	NOS2	10	Text	Gene expression score btw -5 to 10	-2.0000, 1.0000
PD-1	PD1	10	Text	Gene expression score btw -5 to 10	-2.0000, 1.0000
PDL2	PDL2	10	Text	Gene expression score btw -5 to 10	-2.0000, 1.0000
TIGIT	TIGIT	10	Text	Gene expression score btw -5 to 10	-2.0000, 1.0000
Sample Type	SPEC	20	Text	Sample type	DNA
Comments	COM	200	Text	Comments	XXXX

Map: These data here from the Table 5 based DTP represent IO 360™ panel data. The gene expression is measured tumors using the PanCancer IO 360™ panel to provide a comprehensive evaluation of immunotherapy on anti-tumor immune response. IO 360™ panel is a 770 gene expression panel that includes 48 distinct gene signatures to examine cancer related immune responses.

With variety of panels available from IO 360™ data, the DTP under table 5 is designed to collect a list of gene signatures corresponding to 'Inhibitory Immune Signaling' sub panel. These signatures measure the gene expressions that impact the T cells responses by preventing proliferation, antigen presentation, T cell activation, cytokine production. This data aligns with the definition of SDTM IG's definition for Genomics Findings domain (GF).

Considering the assumptions, below is the high level mapping of this example external raw data to GF domain.

- GFCAT here is 'INHIBITORY IMMUNE SIGNALING'
- GFTSTDTL here is 'GENE EXPRESSION'
- GFTEST have values 'ARG1', NOS2', 'PD1', 'PDL2', 'TIGIT' etc
- GFMETHOD here is 'PANCANCER IO 360 PANEL'
- GFSPEC here is 'DNA'
- GFORRES/GFSTRESC/GFSTRESN will have values from the signature result fields
- GFDTC will have value from collection date field
- CO domain will capture the comments.
- Any additional content not part of parent domain will be submitted under SUPP domain.

Cell Phenotype Data:

Cell phenotype data in clinical trials refers to the observable characteristics of a cell collected from a subject during the study, which can include data like characterization of cell phenotype, lineage, and function based on the expression of specific markers in a single cell or cell suspension. The scope of samples include disseminated tissues (blood and other body fluids, bone marrow aspirate) and cell suspensions with cell phenotyping based on the use of markers. This data can be used to understand diseases progression and predict treatment response at cellular level.

Based on the SDTM IG 3.4, this type of phenotype data is recommended to be submitted in a newly introduced Cell Phenotype (CP) domain. The scope of findings include measurement of cell populations identified, classified based on the differential expressions of phenotypic markers, level of marker expression, and other cell properties based on the characterization of expression markers and/or substances interact with markers.

Table 6: Example of a DTP for Phenotype data -

Field Name	SAS Name	Length	Data Type	Explanation	Example
File Creation Date and Time	FILCRDTM	50	Text	The date/time file created by sender	DDMONYYYY XX:XX
External Vendor Name	LABNAM	50	Text	*Vendor name XXX*	
Protocol	STUDNAM	25	Text	Protocol number	
Site ID	SITEID	3	Text	Site number	001, 002, etc.
Subject ID	SUBJID	11	Text	Subject number	001-001-001 etc.
Visit Name	VISIT	40	Text	Visit name explained in DTP	
Collection Date	LBDT	11	Text	Collection date	DDMONYYYY
Visit Type	VISITTYP	1	Text	Planned visit and unscheduled	Day 1, Unscheduled
Accession Number	ACCESNUM	10	Text	Accession number	1234567
Subset Population	SUBPOP	100	Text	Subset cell population	CDxx+CDyy- etc
Gate	GATE	100	Text	Gate definition followed	FSC+SSC+|CD14+ etc
Marker Class	MRKCLASS	100	Text	Marker Classify	ANXV+ etc
Test	TEST	100	Text	Marker used to perform test	CD14+CDxx+CDyy-ANXV+
Sample Type	SPEC	20	Text	Sample type	CSF, Other, Unknown
Method	METHOD	40	Text	Method used to perform analysis	Flow Cytometry
Result	ORRES	40	Text	Quantitative result	1, 2, 3, etc
Unit	ORRESU	10	Text	Result unit	10^6/L, %, blank
Comments	COM	200	Text	Comments	XXXX

Map: These data here from the Table 6 based DTP represent phenotyping data. The testing performed is quantitative cell phenotyping to measure the cell populations based on the expressions of phenotypic markers using flow cytometry. This data aligns with the definition of SDTM IG's definition for Cell Phenotype Findings domain (CP).

Considering the assumptions, below is the high level mapping of this example external raw data to CP domain.

- CPCAT here is 'IMMUNOPHENOTYPING'
- CPTESTs here are 'Mono', 'Mono Sub', 'Lymphocytes', 'Lymphocytes/Leukocytes' etc
- CPSBMRKS here is 'CDxx+CDyy-' based on the subset population
- CPCELSTA here is 'APOPTOTIC' based on the marker class state
- CPCSMRKS here is ANXV+ based on marker class
- CPGATE here is 'MONO' or 'LYM'
- CPGATEDEF here is 'FSC+SSC+|CD14+' based on gate definition
- CPMRKSTR here is CD14+CDxx+CDyy-ANXV+ based on the marker tested
- CPRESSCL here is 'QUANTITATIVE'
- CPRESTYP here is 'NUMBER FRACTION' or 'NUMBER CONCENTRATION'
- CPORRES/CPSTRESC/CPSTRESN will capture the results
- CPORRESU/CPSTRESU will capture the units
- CPMETHOD here is 'FLOW CYTOMETRY'
- CPSPEC here is 'CEREBROSPINAL FLUID'
- CPDTC will capture the Collection Date information
- CO domain will capture the comments
- Any additional content not part of parent domain will be submitted under SUPP domain.

Microbiology Data:

Microbiology data in clinical trials refers to collection and analysis of information like detection, identification, quantification and other characteristics about microorganisms that are non-host infectious like virus, bacteria, fungi

and parasites from subject sample to assess the effectiveness of a treatment, susceptible to treatment or any additional side effects.

Based on SDTM IG 3.4, this type of microbiology findings will be captured under Microbiology Findings (MB) or Microbiology Susceptibility (MS) domain.

Table 7: Example of a DTP for Microbiology data - Detection of Replication competent retrovirus (RCR)

Field Name	SAS Name	Length	Data Type	Explanation	Example
File Creation Date and Time	FILCRDTM	50	Text	The date/time file created by sender	DDMONYYYY XX:XX
External Vendor Name	LABNAM	50	Text	*Vendor name XXX*	
Protocol	STUDNAM	25	Text	Protocol number	
Site ID	SITEID	3	Text	Site number	001, 002, etc.
Subject ID	SUBJID	11	Text	Subject number	001-001-001 etc.
Visit Name	VISIT	40	Text	Visit name explained in DTP	
Collection Date	LBDT	11	Text	Collection date	DDMONYYYY
Visit Type	VISITTYP	1	Text	Planned visit and unscheduled	Day 1, Unscheduled
Result	ORRES	40	Text	RCR presence result	Detected, Undetected, Numeric value
Assay Type	ASSAY	100	Text	Assay used to detect	qPCR
Accession Number	ACCESNUM	10	Text	Accession number	1234567
Sample Type	SPEC	20	Text	Sample type	PBMC, Other, Unknown
Comments	COM	200	Text	Comments	XXXX

Map: RCR is a safety concern where a retroviral vector used to modify the cells in gene therapy could regain the ability to replicate itself and potentially leading to uncontrolled virus spread within patient if present in the administered cell product. FDA recommends to perform RCR testing to detect any traces of RCR present in cell products before administration and as well in the patient samples as a follow up after the cell product is administered. RCR detection assay are used to detect the presence of RCR in viral vectors. The data in the table 7 above is based on the extended and quantitative polymerase chain reaction(qPCR) assay to detect presence of RCR in viral vectors.

This data aligns with the definition of SDTM IG's definition for Microscopic Findings domain (MB).

Considering the assumptions, below is the high level mapping of this example external raw data to MB domain

- MBTESTCD is 'RCR'
- MBTEST is 'Replication Competent Retrovirus'
- MBTSTDTL values are 'DETECTION', 'CELL COUNT'
- MBORRES/MBSTRESC will include values like 'PRESENT'
- MBMETHOD is 'QUANTITATIVE POLYMERASE CHAIN REACTION'
- MBSPEC is 'PEROPHERAL BLOOD MONONUCLEAR CELL'
- MBDTC will capture the value from collection date field
- CO domain will capture the comments
- Any additional content not part of parent domain will be submitted under SUPP domain.

Quality Control:

- **Data transfer as per DTP**: DTP includes the details about variable level metadata, value level metadata, scope of protocol planned assessments/visits to collect subject sample that are sent to a lab to analyze, codelists, CDASH/CDISC compliance, scope of the tests performed, methods used, any exceptions for

sample not analyzed or analyzed result not provided under comments etc. Data import programs, and SDTM programs are setup to process above listed data points based on the approved DTP and test transfer. It is essential to receive the data transfers as per DTP. A metadata level comparison and edit checks are planned for every incoming data transfer from a lab to accept a transfer or reject it due to any discrepancy or deviation from the agreed DTP. Working with labs to align and process the transfer as per DTP will help to maintain a consistent data transfer quality and save time and resources by minimize any downstream impacts of this data. After a DTP is approved, and for any modifications needed, this will prompt for a revision to DTP.

- **Reconciliation:** CRF data (electronic data capture – EDC) will capture the collection/assessment/visit date and time when the sample was collected from a subject. This information is forwarded to a lab who are contracted to analyze the samples. It is important to verify, confirm and reconcile the collection/assessment/visit date and time of analyzed data from a lab with the CRF data. As a part of reconciliation, in a case one could observe collection dates and time discrepancies between the two sources (CRF vs External data). For example, collection date and time present in external data and not present in CRF data or vice versa, or a possible data entry error. It is essential to work with external lab and study site to resolve these reconciliation discrepancies respectively. A clean reconciliation report will help maintain the quality control and eliminate any missing data.

- **Data issues:** Documenting the identified incomplete data points, out of scope data based on DTP, missing data, and non-printable non-ascii characters with the edit check programming, and have those timely resolved with lab will help maintain the quality control.

- **CDISC Compliance:** Designing the DTP to a near CDASH and CDISC standards to minimize the downstream mapping needs at SDTM and ADaM level. Validating the SDTM and ADaM that report external data with pinnacle validator and with in-house programmed compliance checks will help maintain the quality control.

Track:

The scope of a master external data tracker would include information about the list of studies, contracted labs, study needed external data, status of DTP review, CRF forms to perform date reconciliation of external data, targeted SDTM domain, targeted ADaM domain, and comments to document any appropriate additional content. It is essential to maintain this tracker up to date, as this will act as a library of reference while reviewing any new DTPs, and handling and mapping of the external data to save programming teams a quite a set of resources and time. When the input information and source files are consistent, this tracker could be programmatically refreshed to make it up to date at a scheduled frequency. This remain a cross functional (Data management, clinical programming, statistical programming and biostatisticians) tracker to utilize it for the processing of external data. The mapping details and assumptions used here should remain compliance with CDISC standards.

Table 8: Example of a Master External Data Tracker

Row	Protocol	Lab	Ext.Data	DTP Review	CRF Reconciliation	SDTM IG 3.4	ADaM IG 1.2	Comment
1	Study 1	A	PCR	Done	PK collection	PC	ADPC	
2	Study 1	B	ADA	Done	Central collection	IS	ADIS	
3	Study 1	C	Cell of Origin	Done	Tumor Biopsy	GF		Listing use SDTM as source
4	Study 1	D	IO_360	Ongoing	Tumor Biopsy	GF		Exploratory
5	Study 1	E	Phenotype	Done	CSF collection	CP	ADCP	
6	Study 1	F	RCR	Draft	PBMC collection	MB	ADMB	
7	Study 2	B	ADA	Ongoing	Central collection	IS	ADIS	
8	Study 2	H	PCR	Ongoing	PK collection	PC	ADPC	Lab ‘A’ contract ended.
9	Study 3	G	IHC	Done	Tumor Biopsy	MI	ADMI	
10	Study 3	B	ADA	Done	Central collection	IS	ADIS	

Row	Protocol	Lab	Ext.Data	DTP Review	CRF Reconciliation	SDTM IG 3.4	ADaM IG 1.2	Comment
11	Study 4	F	RCR	Done	PBMC collection	MB	ADMB	
12	Study 4	E	Phenotype	Done	CSF collection	CP	ADCP	

Row 1: Shows the study 1 based PCR data from lab A with DTP review complete and is mapped to SDTM PC domain and ADaM ADPC domain with reconciliation performed using the CRF form PK collection using the date and time recorded during assessment.

Row 3: Shows the study 1 based cell of origin data from lab C with DTP review complete and is mapped to SDTM GF. No ADaM domain is planned as the analysis reporting will use SDTM GF as a source to generate a listing.

Row 4: Shows the study 1 based gene expression based on IO 360™ data with DTP review ongoing will be mapped to SDTM GF domain and no ADaM is planned. This data is requested for exploratory purpose only and is documented under comments column.

Row 8: Shows the study 2 based PCR data from lab H with DTP review ongoing with be mapped to SDTM PC domain and ADaM ADPC domain. This PCR data is from a different lab compared to previous studies which were contracted with lab A. While drafting a DTP with new lab H, the previously approved DTP from lab A will be a reference to maintain consistency and for efficient downstream processing of this data. Comment column can includes the additional details.

Visit Mapping of External Data:

Schedule of assessments from the protocol will provide visit information about sample collection. Additionally, based on the therapeutic area of interest, a sample collection at times is subject's event driven. These sample collections could align with the scheduled visit window or could be an unscheduled assessment/visit.

As the EDC recorded assessment/collection date and time, and visit being the primary source of information, the recommended practice is to consider the EDC recorded visit as a source for SDTM visit. This can be processed by merging the visit of EDC with external data with respect to subject, sample, and collection date and time. This corresponding EDC visit will then be post processed using Trial Visits (TV). Additionally, any unscheduled visits from external data will be handled by slotting across the subjects' scheduled visits following the guidance provided in the SDTM implementation guide (IG).

Conclusion

We have provided in this paper our insights and experiences working on external data. We hope these practices and strategies can be utilized and adopted by users to successfully handle and map external data by gathering the information from sources available, understanding the scope of data and downstream analysis needs, designing the DTP, processing and mapping the data to CDISC standards, and effectively track the external data to achieve consistency and standardization. Kindly contact authors with any questions related to this paper.

References

- CDISC SDTM IG V3.4
- FDA Study Data Standards
- Nanostring IO360
- Labcorp IO360
- Cell Phenotyping Data
- Genomics Data
- Microbiology Data
- Biomarker Data
- Pharmacokinetic Data
- PCR

- Flow Cytometry
- Immunogenicity Data
- B-cell Aplasia

Acknowledgements

The authors would like to thank and acknowledge Abhishake Singh and Annette Bove for their review and comments. Thanks to our families for their continuous support.

Contact Information

Your feedback and questions are valued and welcomed. Kindly contact authors at:

Vamsi Kandimalla, M.S
Kite Pharma Inc, a Gilead Company
vkandimalla@kitepharma.com

Murali Kanakenahalli, M.S
Kite Pharma Inc, a Gilead Company
mkanakenahalli@kitepharma.com