Data Standards for Non-interventional Studies

phuse.eu

Contents

Doc ID: WP-029 | Version: 1.0 | Working Group: Optimizing the Use of Data Standards | Date: 15-Sep-2020

Overview: Purpose of this document

Since non-interventional study data do not necessarily need to be submitted to a regulatory agency, such as the FDA, there is no strict requirement for data standards. However, given the increasing interest in RWE, this is likely to change in the future. The purpose of this document is to identify the most common data standards challenges programmers experience while working on non-interventional studies and to suggest the means to deal with these challenges. Feedback from those within various areas of the industry identified the practice of mapping data to SDTM or ADaM conventions as a common issue. The general idea is to stay as close as possible to ADaM with no or minimum modifications.

Scope

Although non-interventional studies are not like the studies most of us come across every day and have their own specifics, they are still similar to the formal clinical trials in some aspects. Of course, each non-interventional study is different and has its own challenges but some of them are more likely to occur. Since non-interventional study data do not necessarily need to be submitted to a regulatory agency, such as the FDA, there is no strict requirement for data standards. However, as ADaM has proved to be incredibly helpful when it comes to the analysis, we will provide information about SDTM and ADaM concepts that can be applied, mostly concentrating on ADaM.

Acronyms

ACDM: Analysis Common Data Model
ADaM: Analysis Data Model
ADY: Analysis Day
AE: Adverse Event
ATC: Anatomical Therapeutic Chemical Classification System
AWHI: Analysis Window Ending Timepoint
AWLO: Analysis Window Beginning Timepoint
AWTARGET: Analysis Window Target
AWTDIFF: Analysis Window Diff from Target
AVISIT: Analysis Visit
BDS: Basic Data Structure
CDISC: Clinical Data Interchange Standards Consortium
CRF: Case Report Form
DTYPE: Derivation Type
EHR: Electronic Health Record
FDA: Food and Drug Administration
FHIR: Fast Healthcare Interoperability Resources
MA: Marketing Authorisation
MAH: Marketing Authorisation Holder
NCBI: National Center for Biotechnology Information
SAP: Statistical Analysis Plan
SDTM: Study Data Tabulation Model (CDISC)
SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms

Definitions

Define-XML is required by the United States Food and Drug Administration (FDA) and the Japanese Pharmaceuticals and Medical Devices Agency (PMDA) for every study in each electronic submission to inform the regulators which datasets, variables, controlled terms, and other specified metadata were used [1].

Non-interventional study: A study where the medicinal product(s) is (are) prescribed in the usual manner in accordance with the terms of the marketing authorisation. The assignment of the patient to a particular therapeutic strategy is not decided in advance by a trial protocol but falls within current practice, and the prescription of the medicine is clearly separated from the decision to include the patient in the study. No additional diagnostic or monitoring procedures shall be applied to the patients, and epidemiological methods shall be used for the analysis of collected data [2]. In this context it is considered important to clarify that interviews, questionnaires and blood samples may be considered normal clinical practice.

Start of data collection: the date from which information on the first study subject is first recorded in the study dataset or, in the case of secondary use of data, the date from which data extraction starts. Simple counts in a database to support the development of the study protocol, for example to inform the sample size and statistical precision of the study, are not part of this definition [3].

End of data collection: the date from which the analytical dataset is completely available [3].

Problem Statement

Currently there are no defined data standards or guidelines for non-interventional studies. In this project we would like to:

- study the challenges that programmers face when creating analysis datasets such as
  - visit handling
  - creation of treatment section
  - performing imputations
  - dealing with questionnaires
  - specific-level data handling
- discuss and optimise the ways these issues can be resolved from programming and data creation perspectives
- describe how to ensure data transparency and traceability.

Background

During this project, statistical programmers were contacted and asked to complete a survey, which asked them to share the challenges faced while working on non-interventional studies:

Q1. What problems or challenges have you faced while implementing SDTM for non-interventional studies?
Q2. What problems or challenges have you faced while implementing ADaM for non-interventional studies?
Q3. What problems or challenges from a CDISC standards

perspective have you faced while mapping data from non-interventional studies that contained missing or incomplete values?

Based on their feedback, and on our own experience, we identified the most common challenges. Almost all the responders noted the lack of standards as the biggest challenge. The table below presents other challenges identified.

Challenge	Number of answers
Impossible to create certain SDTM domains or required variables	7
Incomplete, inconsistent data	3
Other	2

Table 1: Quiz results

Types of Non-interventional Studies

Example 1: Intensive monitoring schemes: Intensive monitoring is a system of record collection in designated areas, e.g. hospital units, or by specific healthcare professionals in community practice. In such case, the data collection may be undertaken by monitors who attend ward rounds, where they gather information concerning undesirable or unintended events thought by the attending physician to be (potentially) causally related to the medication [3].

Example 2: Prescription event monitoring: In prescription event monitoring (PEM), patients may be identified from electronic prescription data or automated health insurance claims. A follow-up questionnaire can then be sent to each prescribing physician or patient at pre-specified intervals to obtain outcome information. Information on patient demographics, indication for treatment, duration of therapy (including start date), dosage, clinical events and reasons for discontinuation can be included in the questionnaire. PEM tends to be used as a method to study safety just after product launch. Limitations of prescription event monitoring include substantial loss to follow-up, relatively short duration of follow-up, selective sampling, selective reporting and limited scope to study products which are used exclusively in hospitals. However, in PEM, there is the opportunity to collect more detailed information on adverse events from a large number of physicians and/or patients [3].

Example 3: Patient registries: Patient registries have been defined as 'an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves a predetermined scientific, clinical, or policy purpose(s). [4]' In this case, data collection is prospective over time and independent of the use of medicine according to the marketing authorisation. There is no protocol-defined treatment, management, or allocation of patients and patient visits. As inclusion in the registry does not have any impact on the therapeutic strategy, such studies are also non-interventional. Patient registries may be classified as:

- disease or condition registries, based on populations with a particular disease or group of diseases (ISPOR definition)
- product-specific registries, based on populations using specific products (often developed by manufacturers to assess long-term safety and adverse effects).

Example 4: Case control studies: Porta's Dictionary of Epidemiology defines the case-control study as an observational epidemiological study of persons with the disease (or another outcome variable) of interest and a suitable control group of persons without the disease (comparison group, reference group) [5]. The potential relationship of a suspected risk factor or an attribute to the disease is examined by comparing the diseased and non-diseased subjects with regard to how frequently the factor or attribute is present (or, if quantitative, the levels of the attribute) in each of the groups (diseased and non-diseased). In case-control studies, the researcher retrospectively reviews the events, including treatment. These studies have no experimental element and their purpose is to examine the events and factors while having no impact on them.

Example 5: Cohort study: A cohort study is a study in which subjects with a certain condition and/or who receive a particular treatment are followed over time and compared with another group of subjects who are not affected by the condition under investigation. For research purposes, a cohort is any group of individuals who are linked in some way or who have experienced the same significant life event within a given period [6]. Both prospective and retrospective data collection are possible. Prospective data collection is non-interventional if assignment to a treatment arm is not decided in advance by a trial protocol.

Example 6: Cross-sectional study: Data collected on a population of patients at a single point in time (or interval of time) regardless of exposure or disease status constitute a cross-sectional study. These types of studies are primarily used to gather data for surveys or for ecological analyses.

Example 7: Case-only designs: Case-only designs have been proposed to assess the association between intermittent exposures and short-term events, including the self-controlled case-series, the case-crossover and the case-time-control studies. In these designs, only cases are used and the control information is obtained from person-time experience of the cases themselves.

Example 8: Safety review of class of medicines: One product has a specific requirement for regular diagnostic testing to monitor adverse events (AEs) while other products of the same class do not. The Marketing Authorisation Holder (MAH) has been requested to perform a cohort study to estimate the real-life incidence of the AEs across the whole class of products. The proposed diagnostic/monitoring procedure could be considered current clinical practice, although it is specified for only one product of the class, and therefore classified as non-interventional [7].

Example 9: Drug utilisation studies: Observing the use of a drug in real life (as opposed to the rigid settings of clinical trials). This could include evaluating patterns of use of a medicinal product, such as capturing off-label use, and can even be conducted with this specific aim. In this case, research is purely observational as there is no experimentation involved [7].

Example 10: Administrative claims data: Administrative claims databases are built on data extracted from claims submitted by healthcare providers to payers when a patient uses health services. They include pharmacy databases and health insurance records and are used to monitor health and disease management. These systems are primarily maintained for

billing and administration, but can also be used by researchers, insurers, health authorities and other stakeholders to provide long-term data on the impact of health interventions on the healthcare system in 'real-world' (observational) studies.

Claims databases generally include information on the use of inpatient, outpatient, emergency room and pharmacy services. They contain information on, for example, the services performed during a clinician's visit, surgical interventions, diagnostics, laboratory tests, hospitalisation and length of stay, and pharmacy filing [9].

Visit Handling

The CRF field "VISIT", common for clinical trials, is frequently not included in the CRF for non-interventional studies. There might be no CRF at all. However, if the endpoints are defined based on certain timepoints throughout the study, the values collected should be somehow assigned to these timepoints.

If the full dates are available in most of the cases, it should still be possible to derive analysis day or ADY, the study day relative to a reference date. It is useful to keep in mind that the study day does not have to be based on treatment start date. According to the ADaM implementation guide [8]: 'The reference date should be indicated in the variable-level metadata for ADY and the reference date should be included as a variable in the given analysis dataset or alternatively in ADSL.' This can be particularly helpful when there is no study drug.

Also, for the timepoints of interest, the programmers are usually able to define a target day or target window. In ADaM terminology these would be stored in AWTARGET (Analysis Window Target) and/or AWLO (Analysis Window Beginning Timepoint) and AWHI (Analysis Window Ending Timepoint). Another variable that may come in handy during the programming is AWTDIFF (Analysis Window Difference from Target) as it represents the difference from target.

After deriving the variables described above, programmers can proceed to visit mapping. Ideally, it should be clear in the Statistical Analysis Plan (SAP) how this should be done. For example, where the closest/first/last/worst/best value needs to be used, the easiest way would be to set ANL01FL to "Y" on the respective value and use this flag for subsetting during the analysis programming. It may also happen that the average needs to be derived, for example if two values are equally close to the target day. In this case, the best way would be to create a new record and set DTYPE (derivation type) to "AVERAGE". In order to maintain data transparency, we strongly recommend not to delete any records from the dataset but rather use analysis flags and derivation type variable for subsetting. Although the data is not necessarily submitted, it is always easier for both programmers and reviewers if traceability is maintained. The table below illustrates how one can apply the AW variables.

Analysis Time Point (AVISIT)	Target Study Day (AWTARGET)	Time Window (AWRANGE)	Lower Limit Day (AWLO)	Upper Limit Day (AWHI)
Baseline	1	–		
1 week	8	day 2 – day 11	2	11
2 weeks	15	day 12 – day 18	12	18
3 weeks	22	day 19 – day 25	19	25
4 weeks	29	day 26 – day 32	26	32

Table 2: Example of windowing

Missing Value Imputation

Missing value imputation goes along with the visit mapping. It is usually done afterwards once it is clear which values need to be imputed. It may also happen that the values are not missing but due to some medical reason specified in the SAP must be excluded from the analysis. The ADaM implementation guide [7] has clear guidance on what to do in these cases: When an analysis timepoint is missing, it is recommended to create a new row in the analysis dataset to represent the missing timepoint. Derivation type or DTYPE variable should be used to identify these rows: For example, when the last/worst observation needs to be carried forward to replace the missing values, create LOCF/WOCF rows, and identify these by populating the derivation type variable DTYPE with values LOCF or WOCF. The original records should have DTYPE populated as null. It would enable the programmers to select the appropriate rows for analysis by selecting DTYPE = null for Data as Observed (DAO) analysis, DTYPE = (null or LOCF) for LOCF analysis, and DTYPE = (null or WOCF) for WOCF analysis.

The following example illustrates this scenario. There is a raw dataset with laboratory data without any visits. For better understanding, it is considered to have SDTM-like variable names.

Row	USUBJID	LBTEST	LBDT	LBORRES	LBORRESU
1	0001	Platelets	2017-02-02	305	(10^9/L)
2	0001	Platelets	2017-02-12	274	(10^9/L)
3	0001	Platelets	2017-02-16	300	(10^9/L)
4	0001	Platelets	2017-02-24	276	(10^9/L)
5	0001	Platelets	2017-02-25	274	(10^9/L)
6	0001	Platelets	2017-03-06	321	(10^9/L)
7	0001	Platelets	2017-03-10	325	(10^9/L)
8	0001	Platelets	2017-03-16	326	(10^9/L)
9	0001	Platelets	2017-03-19	290	(10^9/L)

Table 3: Raw lab values

The analysis is to be done by Baseline, Week 1, Week 2, Week 4, Week 8 and Week 12 timepoints. All the ranges, target dates and rules to apply are usually specified in the corresponding SAP.

Row	AVISIT	AVISITN	AWTARGET	AWRANGE	AWLO	AWHI	AWU
1	Baseline	-1				1	DAYS
2	Week 1	1	8	2–11	2	11	DAYS
3	Week 2	2	15	12–18	12	18	DAYS
4	Week 4	4	29	26–32	26	32	DAYS
5	Week 8	8	57	43–71	43	71	DAYS
6	Week 12	12	85				DAYS

Table 4: AVISIT mapping table from SAP

The first step is to calculate the study day as:

- ADT-Reference Date if ADT<Reference Date
- ADT-Reference Date +1 if ADT>=Reference Date.

As there are no visits, information from Table 4 should be combined with the data of interest (in this case, laboratory – Table 3). One of the ways is to perform a many-to-many merge so each observation in the data of interest would get the whole set of ranges. Then the AVISIT can be assigned. As the ranges should be not overlapping, only one (or none) visit will be assigned to a record. In this example, laboratory values need to be assigned to visits using Table 4. For each visit, values, which fall into the corresponding range, should be considered. If there is more than one value within the window, then the closest value to the target day should be chosen for analysis. If there are multiple results equally close to the target day, then the average is to be derived and used for analysis. If no values fall into the window, the last observation is carried forward.

Row	USUBJID	AVISIT	AVISITN	LBDT	AVAL	ADT	DTYPE	ADY	AWTARGET	AWTDIFF	AWLO	AWHI	ANL01FL
1	0001	Screening	-1	2017-02-02	305	02FEB2017		-5					Y
2	0001	Week 1	1	2017-02-12	274	12FEB2017		6	8	2	2	11	
3	0001	Week 1	1	2017-02-16	300	16FEB2017		10	8	2	2	11	
5	0001	Week 1	1		287		AVERAGE		8		2	11	Y
6	0001	Week 2	2	2017-02-24	276	24FEB2017		18	15	0	12	18	Y
7	0001			2017-02-25	274	25FEB2017		19					
8	0001	Week 4	4	2017-03-06	321	06MAR2017		28	29	1	26	32	Y
9	0001	Week 4	4	2017-03-10	325	10MAR2017		32	29	3	26	32	
10	0001			2017-03-16	326	16MAR2017		38					
11	0001			2017-03-19	290	19MAR2017		41					
12	0001	Week 8	8	2017-03-19	290	29MAR2017	LOCF	41	57	16	43	71	Y
13	0001	Week 12	13	2017-03-19	290	29MAR2017	LOCF	41	85	16	43	71	Y

Table 5: AVISIT Mapping

In this case, the traceability to raw data can be established via USUBJID, LBDT and LBTEST variables. For AVISIT=Week 1, there are two values on study days 6 and 10 which are equally close to the target day. Therefore, an additional row is created with DTYPE=AVERAGE and ANL01FL is set to Y for this derived record. For AVISIT=Week 4, there are also two values on study days 28 and 32, but the one with ADT=06MAR2017 is closer to the AWTARGET and therefore marked with ANL01FL=Y. As there is no information after 19MAR2017, for both AVISIT=Week 8 and AVISIT=12 the LOCF rule is applied and the result from ADT=19MAR2017 is carried forward. LBTEST and LBDT are kept for these derived records to clarify the source.

Treatment Section

Observational study data do not fit into ADaM when participants/patients receive no study drug. According to the ADaM implementation guide [8], ARM and TRT01P are required. However, because the data are not submitted to the FDA or any other regulatory agency it is not crucial to have the treatment section. When it comes to the derivation of study day, a date other than TRTSDT can be used as well, according to the CDISC note from the guide:

- 'The relative day of AVAL and/or AVALC. The number of days from an anchor date (not necessarily DM.RFSTDTC) to ADT.'
- 'Note that it is possible to have different definitions for a relative day (or time) variable (e.g. ADY) in separate datasets, using different anchor dates (or times). For example, the derivation of ADY for efficacy datasets might be different from that for safety datasets.'

If programmers want to have ARM and TRT01P, one of the options would be to populate dummy values such as "No treatment" or "No drug". For cohort studies, it can make sense to store the cohort information in the treatment variables. Although it is not treatment information, the analysis in such studies is often performed by cohort, similarly to by-treatment analysis in interventional studies. In the situation where one subject can be included in multiple cohorts, a separate set of COHORT variables – e.g. COHORT01, COHORT02 – can be introduced, similar to TRTXXP/TRTXXPN, which would capture all the cohorts the subject is included in. If the date when the subject entered the cohort is of interest, cohort entry date (and time) may also be stored in a numbered variable following the same naming convention. These variable names should end with "SDTM" for datetime and with "SDT" for date only, which state for start datetime and start date, respectively.

Date Imputation

Working on non-interventional studies usually involves quite a lot of date imputation. The information is often collected retrospectively. For many dates, the day and/or month are missing; however, these data must still be analysed. In this section, we look closer at date imputation. From experience,

the worst-case scenario is the best one when it comes to date imputation. For example, if an adverse event could have started before or after initiation of the study drug, the event is considered to have started after the treatment start date and therefore treatment-emergent. ADaM [7] has a set of variables that are used if date imputation is required. These are *DTF and *TMF variables: '*DTF variables represent the level of imputation of the *DT variable based on the source variable. *DTF = Y if the entire date is imputed. *DTF = M if month and day are imputed. *DTF = D if only day is imputed. *DTF = null if *DT equals the respective variable date part equivalent. If a date was imputed, *DTF must be populated and is required. Both *DTF and *TMF may be needed to describe the level of imputation in *DTM if imputation was done.'

Several survey responses indicated data collected in different formats are a big problem. This can be handled quite successfully using the ANYDTDTEw. When programming is done in SAS, there is an informat which can work with the dates in the following formats: DATE, DATETIME, DDMMYY, JULIAN, MDYAMPM, MMDDYY, MMxYY<YY>, MONYY, TIME, YMDDTTM, YYMMDD, YYQ, etc. In order to avoid potential issues with ambiguous dates, programmers can use the DATESTYLE option, which indicates the order of month, day and year. Based on the data source location they can then set up the option to the respective value.

Interim Analyses

For observational studies, it is often required to perform interim analyses. For example, it may be envisaged in the SAP to do such analyses every year using cutoff dates. Cutoff of findings data should not be a problem; however, it may be challenging for events and subject status at the cutoff date. For example, if event stop date is after the cutoff date, the event should be considered ongoing.

The cutoff date can be saved in ADSL in EOSDT and the subject status at the cutoff date in EOSSTT. Similarly, the treatment status information may be placed into EOTSTT. To perform cutoff of the data it may be helpful to put the cutoff date into a global macro-variable. This would allow changes to be made to only one macro-variable for following interim analyses. Also, additional endpoints may be applicable for the interim analysis.

Questionnaire Data

Questionnaire data represents quite a significant part of data collected for non-interventional studies. For this kind of data, ADaM principles are applicable as usually this data fit quite well in the ADQS domain. Currently there are three questionnaire supplements in the SDTM IG 3.3 [10]:

- **Questionnaires (QS)** – Questionnaires have a defined standard structure, format, and content; consist of conceptually related items that are typically scored; and have documented methods for administration and analysis. Most often, questionnaires have as their primary purpose the generation of quantitative statistics to assess a qualitative concept.
- **Functional Test (FT)** – A functional test is an objective measurement of the performance of the task by the subject in a specific instance. Functional tests have documented methods for administration and analysis and require a subject to perform specific activities that are evaluated and recorded. Most often, functional tests are direct quantitative measurements.
- **Clinical Classifications** – If the instrument is a Rating or Grading Scale in which the intent of the instrument is to evaluate a single body system, it would be stored in the morphology/physiology domain which represents that body system. Other Rating or Grading Scales related to multiple body systems and all Composite Score type instruments would be represented as a Clinical Classification in the RS domain.

All these data can be included in the ADQS dataset.

NCBI defines 4 types of clinical outcomes assessments, which are classified into respective questionnaire supplements [11]:

Outcome	Definition	Supplement	ADaM Domain
Clinician reported outcome (ClinRO)	A ClinRO is based on a report that comes from a trained healthcare professional after observation of a patient's health condition. A ClinRO measure involves a clinical judgement or interpretation of the observable signs, behaviours, or other physical manifestations thought to be related to a disease or condition. ClinRO measures cannot directly assess symptoms that are known only to the patient (e.g. pain intensity).	Questionnaire Clinical Classification	ADQS
Observer Patient reported outcome (ObsRO)	An ObsRO is a measurement based on an observation by someone other than the patient or a health professional. This may be a parent, spouse, or other non-clinical caregiver who can regularly observe and report on a specific aspect of the patient's health. An ObsRO measure does not include medical judgement or interpretation.	Questionnaire Clinical Classification	
Patient reported outcome (PRO)	A PRO is a measurement based on a report that comes from the patient (i.e. study subject) about the status of the patient's health condition without amendment or interpretation of the patient's report by a clinician or anyone else. A PRO can be measured by self-report or interview, provided the interviewer records only the patient's response.	Questionnaire	
Performance outcome (PerfO)	A PerfO is a measurement based on a task(s) performed by a patient according to instructions administered by a healthcare professional. Performance outcomes require patient cooperation and motivation.	Functional Test	

Table 6: Clinical outcomes assessments

It is recommended to store all these data in ADQS. In general, the considerations for creating ADQS are in alignment with general basic data structure or BDS rules:

- Dataset contains one or more records per subject, per analysis parameter, per analysis timepoint.
- Question description is put into PARAM variable.
- Question short name is put into PARAMCD variable.
- Answer or result is put into AVAL/AVALC variable.
- If needed, the questions can be grouped using PARCAT1 and PARCAT2 variables.
- It is recommended that PARAMCD for total scores should provide traceability to the individual questions.

While working on non-interventional studies, programmers may come across questionnaires including data that would be more accurately stored in some other domain. For example, demographic characteristics may be collected with other data in a questionnaire. In this case, it might be more suitable to store these data elsewhere, i.e. in ADSL, as this is where most programmers would expect to find it.

Specific-level Data and Matched Studies

It may not be very easy to determine what should be stored in the USUBJID variable. For example, in situations where both observer and patient are of interest, it may be difficult to establish which of them should be considered the subject. In this case, a solution is provided by CDISC. It is possible to have two sets of person-data: one for patient and one for the associated person (observer, investigator, donor, etc.). CDISC provides a set of SDTM associated persons domains which are similar to the usual SDTM domains by their structure and can therefore be transformed into ADaM using the same techniques. Currently, the following example domains are represented in the guide on associated persons [12]: APDM, APEX, APSU, APAE, APMH, APLB, APQS, APRP and APSC. These datasets can be used as the basis for respective ADaMs, e.g. APADSL, APADEX, APADAE. The main difference to the 'usual' domains are the following variables:

Outcome	Definition	Supplement
APID	Associated Persons Identifier	associated persons, or a pool of associated persons. If APID identifies a pool, POOLDEF records must exist for each associated person.
RSUBJID	Related Subject	Identifier for a related study subject or pool of study subjects. The subject(s) may be human or animal. RSUBJID will be populated with the USUBJID of the related subject or the POOLID of the related pool. RSUBJID will be null for data about associated persons who are related to the study but not to any of the study subjects.
RDEVID	Related Device	Identifier for a related device. RDEVID will be populated with the SPDEVID of the related device.
SREL	Subject, Device, or Study Relationship	If RSUBJID is populated, SREL describes the relationship of the associated person(s) identified in APID to the subject or pool identified in RSUBJID. If RDEVID is populated, SREL describes the relationship of the associated person(s) identified in APID to the device identified in RDEVID. If RSUBJID and RDEVID are null, SREL describes the relationship of the associated person(s) identified in APID to the study identified in STUDYID.

Table 7: Specific-level data variables

These variables can be easily carried forward to ADaM datasets. For SDTM, when a person can have multiple relations to the subject, the APRELSUB dataset is introduced. This dataset stores information about all the relations (or only multiple ones) in the structure: one relation per person per subject. If there is a need for creating the person-level ADaM dataset for associated persons (APADSL), it may be necessary to either create multiple relation variables – e.g. SREL1, SREL2 – or concatenate all the values in one variable using a delimiter. The situation when one person can be related to multiple subjects is not covered in the guide, but it is also possible and can be handled in the same way.

Matching is a statistical technique, which is used to evaluate the effect of a treatment by comparing the treated and the non-treated subject in a non-interventional study. The goal of matching is, for every treated subject, to find one (or more) non-treated subject(s) with similar observable characteristics against whom the effect of the treatment can be assessed. The above concept can also be applied for matched studies when APID and RSUBJID are used to capture the subject and matched subject relationship.

Hands-on Experience

While working on our project we gathered feedback from programmers who have actually tried to implement mapping of non-interventional study data to ADaM. For example, in the paper [13] presented at the PHUSE EU Connect 2019, the authors shared their experience of mapping the data needed for their analyses from FHIR to ADaM: The authors managed to map all variables that were required for their analyses. Several issues required special handling. In order to create the "treatment" variable, they assigned subjects who were taking Metformin Hydrochloride at their first encounter in the database to the 'Metformin Hydrochloride' treatment group. Otherwise, the subjects were assigned to the 'Comparator' treatment group. Treatment group was assumed to remain constant for subsequent subject encounters. The authors used patient encounter number as a proxy for visit number. In contrast to a research protocol, patient encounters did not occur at regular pre-defined intervals and the time between subject encounters differed for each subject. Reported AE term was coded using SNOMEDCT and the authors did not recode it in order to be consistent with most EHR systems. This highlights the issue that in many cases there is a lack of harmonisation between EHR/FHIR and CDISC/research data. The authors of the paper emphasised the lack of standardisation between FHIR and CDISC as the main challenge they encountered while mapping their data. Different controlled terminologies used by FHIR and CDISC also complicate the mapping and should be taken into account. ADaM variables are tailored for MedDRA and ATC while in many non-interventional studies different dictionaries and classification systems are used. In this case, a separate set of variables consistent with the respective coding systems needs to replace "classic" ADaM variables.

In some companies the programmers use their own data models to analyse the data for non-interventional studies. In the scope of our project we studied one, called ACDM. The model combines features of ADaM and SDTM. It has similar classes of domains to CDISC SDTM, i.e. interventions, findings, events and special purpose domains. Unlike ADaM, ACDM is designed specifically for non-interventional studies and takes into account some of their specifics like dictionaries and coding systems. On the other hand, while datasets contain some derived variables, they still require quite a lot of post-processing in order to produce the outputs. After a discussion with the programmers involved in this project the conclusion is that ACDM can possibly be mapped to ADaM with several modifications. Many of them, for example, need variables to store matched subjects, which are addressed in this white paper.

Data Transparency

Data transparency is one of the key aspects of every mapping process. Reviewers should be able to trace the data from the analysis dataset back to the raw data and clearly see how the derivations were made. To support this, we recommend keeping original records and making use of ADaM variables such as DTYPE and ANLXXFL as much as possible. This will allow reviewers to distinguish between original and derived records. Another important aspect is keeping original variables, e.g. for dates, and creating additional variables with the imputed values and information about the imputation. Using the SDTM naming conventions for the variables might be an option as SDTM standards are well known in the industry and it is allowed to have them in ADaM as well. While applying the visit windowing it is better to keep all the records in the data with either missing AVISIT or AVISIT assigned to "Unscheduled". This way it will be possible to follow the mapping algorithm and explain which records were mapped and how. Although there are no strict requirements for dataset specifications for non-interventional studies, maintaining proper documentation is also a vital component of data transparency. Creating a proper Define-XML document is not mandatory in this case; however, having a document of similar structure is extremely helpful especially for studies that last several years. It is recommended to have the following:

- general rules and conventions, e.g. date formats, lengths of standard variable
- dataset metadata: description (a short description of the type of information), structure (the level of detail represented by individual records in the dataset), key variables (variables used to uniquely identify and index each record in a dataset)

Name	Description	Structure	Keys
ADSL	Subject-Level Analysis Dataset	One level per subject	USUBJID
ADVS	Vital Signs Analysis	One level per subject per visit per parameter	USUBJID, AVISIT, PARAMCD

Table 8: Dataset metadata example

- variable metadata: name, label, type, length, derivation method, controlled terms or format

Name	Label	Type	Length	Derivation	Controlled terms/format
ANL01FL	Analysis Flag 01	Char	1	Set to “Y” if ...	Y, null

Table 9: Variable metadata example

- value-level metadata: derivations of values for separate parameters

PARAMCD	PARAM	AVAL
PARAM1	Parameter 1	Set to AVAL (where PARAMCD=”PARAM2”) +AVAL (where PARAMCD=”PARAM3”)

Table 10: Value-level metadata example

Additional variables can be added to the datasets in order to improve transparency and facilitate assessment of validity. In the scope of our project, we reviewed the paper on how to improve reproducibility and facilitate validity assessment for healthcare database studies [14] and made a short summary of variables that programmers can add into their ADaM database in order to assist reviewers.

Parameter/Variable	Description	Example, if needed	Core
Data extraction date (DED)	The date (or version number) when data were extracted from the dynamic raw transactional data stream (e.g. date that the data were cut for research use by the vendor).	The source data for this research study was cut by [data vendor] on 1st January, 2017. The study included administrative claims from 1st January 2005 to 31st December 2015.	Recommended
Source data range (SDR)	The calendar time range of data used for the study. Note that the implemented study may use only a subset of the available data.		Recommended
Study entry date (SED)	The date(s) when subjects enter the cohort.	We identified the first SED for each patient. Patients were included if all other inclusion/exclusion criteria were met at the first SED. We identified all SEDs for each patient. Patients entered the cohort only once, at the first SED where all other inclusion/exclusion criteria were met. We identified all SEDs for each patient. Patients entered the cohort at every SED where all other inclusion/exclusion criteria were met.	Recommended
Person- or episode-level study entry	The type of entry to the cohort. For example, at the individual level (1x entry only) or at the episode level (multiple entries, each time inclusion/exclusion criteria met).		Optional
Type of exposure	The type of exposure that is captured or measured, e.g. drug versus procedure, new use, incident, prevalent, cumulative, time-varying.		Optional

Table 11: Variables to support data transparency

Additionally, if the data is coming from different medical institutions, it will be useful to add a variable to indicate the source of each record.

Conclusion

Non-interventional studies have some distinctions compared to interventional clinical trials. As opposed to clinical trials, where every aspect is regulated and CDISC standards apply for the whole study conduct from data collection to the analysis itself, programmers who work on non-interventional studies do not have a shared set of standards to reference. Having no well-defined standards or guidelines can lead to confusion and additional effort, especially for programmers who have only worked on regulated interventional clinical trials.

In this white paper, the project team strived to address the challenges faced in relation to non-interventional studies by collecting detailed feedback and experience from programmers and stakeholders. While this white paper mostly concentrated on how ADaM can be used for non-interventional studies, there are many aspects that also contribute to the dataset programming regardless of the chosen standard. Different coding systems for adverse events and concomitant medications and different data sources add an additional level of complexity. These and a potential intermediate step between raw data and ADaM-like data are the topics that could require further investigation. The interest in real-world evidence data is growing and even though currently such data does not have to be submitted to the regulatory agencies, this is likely to change in the future, and the better the companies are prepared for this, the easier it will be to adapt to this change. To sum up, we recommend implementing ADaM standards whenever possible. The challenges identified in the scope of the project can be handled using either ADaM or ADaM-like concepts. Although it may not be worth fully mapping the data to SDTM, this can be done to a certain extent as an intermediate step. The programming team should assess how beneficial it is for further mapping and how SDTM-compliant they plan to be. It is also helpful to consider SDTM concepts which can be further transformed to ADaM-like, where no ADaM concept is available, for example when dealing with specific-level data.

References

[1] Define-XML. *https://www.cdisc.org/standards/data-exchange/define-xml*. (Access date 13 Sep 2020).

[2] EU Directive. (2001). Directive 2001/20/EC Of The European Parliament And Of The Council. (Access date 11 Mar 2019).

[3] Guideline on good pharmacovigilance practices (GVP). EMA/813938/2011 Rev 3*. 9 October 2017.

[4] Registries for Evaluating Patient Outcomes: A User's Guide [Internet]. 3rd edition.

[5] Porta, M. (Ed.). (2008). A Dictionary of Epidemiology (5th ed.). New York: Oxford University Press.

[6] Cohort Studies. Web Center for Social Research Methods. (Access date 11 Mar 2019).

[7] Schnetzler, G., on behalf of the ENCePP Task Force. Interpretation of the definition of non-interventional trials under the current legislative framework ("Clinical Trials Directive" 2001/20/EC)

[8] ADaM Implementation Guide v1.1. *https://www.cdisc.org/standards/foundational/adam*. (Access date 11 Mar 2019).

[9] Real-world evidence (RWE) Navigator. *https://rwe-navigator.eu*. (Access date 13 May 2019).

[10] SDTM Implementation Guide 3.3. *https://www.cdisc.org/standards/foundational/sdtmig*. (Access date 09 Jun 2019).

[11] Cunningham, G., Kopko, S., LaPann, K. Development & Use of Questionnaire Supplements. CDISC Webinar. 4 June, 2015.

[12] Study Data Tabulation Model Implementation Guide: Associated Persons Version 1.0. *https://www.cdisc.org/system/files/all/standard_category/application/pdf/sdtmig_ap_v1.0*. (Access date 09 Jun 2019).

[13] Abolafia, J., Gopal, P., Hume, S. Use of FHIR in Clinical Research: From Electronic Medical Records to Analysis. PHUSE EU Connect 2019.

[14] Wang, S. V., Schneeweiss, S., et al. Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. Pharmacoepidemiol Drug Saf. 2017 Sep;26(9):1018–1032.

Project Contact Information

Yuliia.bahatska@roche.com

Acknowledgements

Wendy Dobson, Lauren White and the PHUSE team

PHUSE DATA STANDARDS FOR NON-INTERVENTIONAL STUDIES PROJECT TEAM: Alex Nasr, Karen Horton, Renate Kroll, Igor Savostianov, Jeremy Teoh, Lovi Sandhu, Murali Neela, Joe Maskell, Sarah Marsall, Sanjeev Kommera, Pavan Tirunahari, Tomoko Sugihara, Jon Neville, Janet Reich, Yogesh Pande, Sarad Nepal, Sangeeta Sama, Jorge Abate, Dorothy Dlamini, Matthew Harrington, Tanvi Gupta