Paper DS11

Data integration on CDISC ADaM Level: Multi-purpose Structures of ADSL and BDS Datasets

Renate Scheiner-Sparna, Alira Health GmbH, Munich, Germany

ABSTRACT

A data integration has to cover different requirements of the sponsor (ISS, ISE, publications, requests from authorities, management decisions). This presentation compares different structures of ADSL and BDS datasets to meet these expectations. ADSL has one line per subject and study. The integrated period can be represented either by integrated variables, or by a separate line, which covers all the single studies. As third possibility, a separate analysis dataset can hold the information.

A use case of a rare disease unit is presented. The identifiers allow for either drawing information on the whole period, or on one or more of the single studies. The entities of the single studies can be used as a whole, or one or more can easily be selected.

The dataset structure will be compared to the ones presented in other publications (Zhong et al. [1]; Phelan et al. [2]; Minjoe [3].

INTRODUCTION

Integration data means to combine data residing in different sources and providing users with a unified view of them. The data being integrated are somehow heterogeneous in structure and are transformed to a single coherent data store (Lenzerini, 2002 [5]) In the context of clinical studies, this means either combining multiple studies, or combining data from different sources like patient files, insurances, electronic health devices. For this paper, we are looking at the first case, combining multiple on the same substance to improve the quality of analysis results by using all information available.

Data integration (DI) requires a huge time investment, as the single studies in most cases have differences in definitions and data capture, which have to be harmonized. Another important aspect is the considerations of transparency and traceability back to the source data, which needs more effort than in a single study. Decisions that go beyond the standards have to be taken.

The programmer should evaluate in advance which questions will be addressed to the data in the future. This corresponds directly to the data structure, as there is a range of freedom for the programmer to decide, within the possible layouts of analysis datasets.

A simple case would be if each subject participates just in one study of the integrated studies. The data structure looks like a single study in the end. It has one line per subject in ADSL. This structure is suitable if the integrated data is used for one specific purpose. In this case, the integration can be tailored exactly to this task. A more complex design is not needed.

There are more complex structures, which at the same time are more flexible to address different needs. They can hold the integrated, as well as the single study information in the ADaM datasets. If you get different requests on the data, e.g. on the whole period of all studies, but also on parts of the period or on subgroups of studies, then a structure like this can be more suitable.

This article will investigate the advantages and challenges of data integration with repeating lines per subject and study in ADSL.

COMPLEX DATA INTEGRATION STANDARDS DEVELOPMENT (2018-2024)

Since 2015, the CDISC analysis data model team developed an ADaM guideline on data integration (ADAMINT). The ADAMINT draft was not available online, so the contents were not generally known. In 2018, the standards for integrated BDS and OCCDS were added to the ADAMINT. It was under review until 2019.

In 2018 and 2019, a few examples of DI were published, which were using ideas of the ADAMINT draft. The papers by Zhong et al. [1] and Phelan et al. [2] served us as best examples

In 2019, the publication of the ADAMINT was cancelled in the last minute, after all the public review comments had been resolved. These facts are presented in a paper from Minjoe (2024). The main reason for not publishing it was a statement from the FDA. The FDA claimed that they would not accept an ADSL structure with more than one line per unique subject, even in complex data integration. Consequently, the ADAMINT was not published. Until today, no CDISC standard on DI is available.

The Minjoe (2024) publication also presents a new idea for a DI ADaM structure which is consistent with the demand of FDA and at the same time with the rules of CDISC. It could facilitate DI and serve as the base idea for a future standard guide for complex DI. Details will be discussed below in a separate section.

DRAFT ADAM INTEGRATION GUIDELINE

As this was an important milestone in the standards development for DI, let's have a closer look to the ADAMINT. The ADAMINT introduces POOL/POOLN variables as identifiers in analysis datasets. They are designed to be used for pooling categories. They allow for more than one record per USUBJID in integrated ADSL (IADSL) and more than one record per USUBJID and PARAM in integrated BDS (IBDS). According to ADAMINT, IADSL has one line per unique subject per pool. BDS datasets have one line per USUBJID per POOL per PARAM. In addition, there is a line for POOL="Overall" of a subject in IADSL.
A pool can be thought of as a cohort. It is defined by an SAP and can be e.g. the type of study "Comparison" for all placebo-controlled studies. If a subject took part in a placebo-controlled study, it gets a row with POOL="Comparison" and one with POOL="Overall". The POOL="Overall" rows are expected to hold information on the entire period the subject attended in any of the studies, e.g. treatment periods. For some standard variables like AGE, special methods are needed to be defined, how they should be filled in DI.

Another feature of the ADAMINT structure must be explained: Variables STUDYID and STUDIES are mandatory in IADSL. Traceability back to the single study works by using the STUDYID variable, it refers us back to the single study the information of the data record comes from. Variable STUDIES holds the information in which other studies the subject took part. Here, all the studies in which the subject took part are listed. This variable is filled for POOL="Overall" lines, whereas variable STUDYID is filled for all lines that don't have POOL="Overall". Consequently, either STUDYID or STUDIES can be filled, not both.

USE CASE

The data integration of this use case was done in a rare disease unit. We integrated data from about 10 single studies on the same rare disease and medicinal product. We followed the path of studies from phase II to IVb. As usual in rare disease research, we had few subjects in the studies. The registry study is the only one that has nearly 100 subjects.
The subjects took part in one or more study (up to 7). In some studies, the inclusion criteria covered only subjects who came from one of the earlier studies.

INTENTION OF THE DATA INTEGRATION

First of all, the DI was an effort to get a transparent view of all the data. The data came from different data models. From some studies, data were available as raw data, some as SDTM data, and some as analysis data.
At the time of DI, there was no specific analysis plan to be followed. The approach was explorative.
The main expectation was that the data could be analyzed quickly for various items. General safety, queries about specific events of interest, descriptive statistics of subjects, pooling of treatment phases to analyze safety parameters were some of the questions raised.

DESCRIPTION OF USE CASE DATA STRUCTURE

At the time of designing the structure for our use case (2023), we knew that there was a CDISC ADAMINT in the review process. We did not know that the publication of ADAMINT had already been cancelled in 2019. All we knew about the contents of the ADAMINT draft came from publications.

After having studied the publications describing the ADAMINT, we chose a complex data structure with more than one line per subject in ADSL. As the study level information was of high importance for understanding the data and using it, we decided to have one line per unique subject and study. We named the analysis datasets IADSL (I for integrated), which is the complex ADSL class name proposed in the ADAMINT. A subject who took part in only one study was represented by one line (USUBJID=001 in Table 1), a subject who took part in three studies by three lines (USUBJID=03-024 in Table 1).
In addition to the STUDYID variable, we use a STUDIES variable which displays all the studies the subject participated in.
As a rule for the integrated USUBJID, we chose the USUBJID of study 03 if the subject participated in it, because this study was the recently closed one and had a high number of participants at the start of integration programming (in the tables below: Study 03). If the subject did not participate in it, the USUBJID of its first study was used.
We implemented POOL variables for each subject for grouping the subjects/studies according to their study design ("Comparison" vs. "Overall").
Tables 1 and 2 show an extract of the variables important to understand the structure. Four subjects are shown, as can be seen in column USUBJID. The first two subjects have just one row, as they participated only in one study, The third subject has two rows in IADSL for participation in two studies and the last one three for three studies.
TRTSDT holds the first treatment date in any study attended. It is equal in all lines for a USUBJID. The corresponding rule is true for TRTEDT, which is not shown here as a column. Variables TRT0xA hold the actual treatment in the single studies in the sequence of the studies attended, e.g. for USUBJID=03-024: TRT01A is the treatment in study 01, TRT02A the one in study 03. TRT0xSDT and TRT0xEDT contain the start and stop dates for the TRT0xA. Only TR01SDT is shown here for reason of space. Integrated analysis periods for treatment over studies are built in APERIODx with start and end dates in APERxSDT, APERxEDT, shown in table 1 for APERIOD1 and APERIOD2. Please note that subject 03-024 has two treatment periods (in studies 01 and 03), which are interrupted by a not treated period. The treatment period from Study 03 would be covered in APERIOD3, which is not displayed here.

Table 1: Extract of IADSL

STUDYID	SUBJID	USUBJID	STUDIES	POOL	POOLC	TRTSDT	TRT01SDT	TRT01A	TRT02A
Study 03	001	03-001	03	1	All verum	10FEB2022	10FEB2022	AAA	
Study 02	002	02-002	02	2	Comparison	06JUN2018	06JUN2018	AAA	
Study 01	002	03-005	01-03	1	All verum	09SEP2021	09SEP2021	AAA	AAA
Study 03	005	03-005	01-03	1	All verum	09SEP2021	02JAN2023	AAA	AAA
Study 01	003	03-024	01-02-03	1	All verum	03APR2019	03APR2019	AAA	NOT TREATED
Study 02	013	03-024	01-02-03	2	Comparison	03APR2019	01DEC2020	AAA	NOT TREATED
Study 03	024	03-024	01-02-03	1	All verum	03APR2019	09MAR2021	AAA	NOT TREATED

TRT03A	APERIOD1	APER1SDT	APER1EDT	APERIOD2	APER2SDT	APER2EDT
	AAA	10FEB2022	03JUN2022			
	AAA	06JUN2018	02NOV2023			
	AAA	09SEP2021	20AUG2023			
	AAA	09SEP2021	20AUG2023			
AAA	AAA	03APR2019	30NOV2020	NOT TREATED	01DEC2020	08MAR2021
AAA	AAA	03APR2019	30NOV2020	NOT TREATED	01DEC2020	08MAR2021
AAA	AAA	03APR2019	30NOV2020	NOT TREATED	01DEC2020	08MAR2021

For the BDS table, we have an example for one parameter in IADQS. Some records for three of the four subjects from the IADSL are shown. For Transparency, variables STUDYID and VISIT trace back to the single study. AVISIT is the integrated visit over the whole period of the subject in all studies.

Table 2: Extract of BDS Table IADQS

STUDYID	SUBJID	USUBJID	POOL	POOLC	TRTSDT	PARAMCD	AVISIT	VISIT	ADT
Study 03	03-001	03-001	1	All verum	10FEB2022	BPI	Baseline	Baseline	09FEB2022
Study 03	03-001	03-001	1	All verum	10FEB2022	BPI	Month 6	Visit 2	12AUG2022
Study 01	01-002	03-005	1	All verum	09SEP2021	BPI	Baseline	Baseline	08SEP2021
Study 01	01-002	03-005	1	All verum	09SEP2021	BPI	Month 3	Visit 1	08Dec2021
Study 03	03-005	03-005	1	All verum	09SEP2021	BPI	Month 16	Baseline	01JAN2023
Study 01	01-003	03-024	1	All verum	03APR2017	BPI	Baseline	Baseline	02APR2019
Study 02	02-013	03-024	2	Comparison	03APR2017	BPI	Month 20	Baseline	30NOV2020
Study 03	03-024	03-024	1	All verum	03APR2017	BPI	Month 21	Visit 1	08JAN2021
Study 03	03-024	03-024	1	All verum	03APR2017	BPI	Month 23	Baseline	08MAR2021

The merge with IADSL depends on the intention. It can be done in two ways.

1. If the interest is in the general integrated period:

We only need one row of each subject with the integrated variables (here: the first treatment period variables), which are equal in all rows. We want to exclude the studies with comparison design. It is recommended to keep only integrated variables and drop all study-specific ones, because they can cause confusion when they are added.

```
proc sort nodupkey data=IADSL out=IADSL_single
                                (keep=USUBJID APERIO1 APER1SDT APER1EDT);
  by USUBJID;
run;
```

```
data IADQS_merged;
  merge IADQS (in=inqs) IADSL_single;
  by USUBJID;
  if inqs and POOL=1 then output;
run;
```

2. If the interest is in one or more of the single studies:

We need all the rows of the studies of interest, Study 03 in our example. For this task, we need to include the STUDYID identifier.

```
data IADQS_merged;
  merge IADQS (in=INQS) IADSL;
  by USUBJID STUDYID;
  if INQS and POOL=1 then output;
run;
```

COMPARISON OF USE CASE TO DRAFT ADAM INTEGRATION GUIDELINE

Now we will compare the data structure of our use case to the ADAMINT.
There are three main differences compared to this data structure:

1. Variables POOL and STUDYID are used for grouping.

In our use case, a POOL variable is used for the treatment setting in a study (“Comparison” vs. “All verum”). Like in ADAMINT draft, we have one record per subject per pool.
In addition, and more important for our project, we use variable STUDYID for grouping the data. It allows for a high and intuitive traceability. Regarded from the functional aspect, STUDYID is used like a POOL variable. If we would have renamed it to POOL, a new set of records would have been necessary.
To be concise, we have two grouping variables, STUDYID and POOL, in horizontal rather than vertical direction.

2. Records for POOL=”Overall” are not derived.

At the start of integration, the extra line was not regarded to be necessary. The main focus was on the grouping of single studies. When it came to questions on the whole period, information was stored in integrated variables. As an example, general treatment start was stored in ITRTSDT, defined as the first use of the study drug in any study. There are more complex examples, like some important characteristics that are not available in all studies, but only in some. These integrated variables have the same values in all the study lines.
There is always the possibility to store integrated information in two ways, either in extra variables, or in a separate data line. The longer we are working on the DI, the more integrated variables are needed. The IADSL gets very broad, with currently 115 variables. From this experience, a separate “Overall” line for integrated information can be recommended. In case the “Overall” line is not used but integrated information is stored in additional variables, a naming convention for the integrated variables is highly recommended (e.g. TRTSDT - ITRTSDT). Even if the ADaM specifications hold the correct definition how to integrate the data for a variable, it is helpful to see on the first glance which are the integrated variables. A “I” as first letter is a good solution, e.g. to store the date of first active treatment in ITRTSDT and the treatment start in a single study in TRTSDT. In most cases the standard variable name has not more than 7 characters, so the I can be prepended without cutting at the end.

3. Records Variables STUDYID and STUDIES are filled both in each line.

As explained above, STUDYID is used like a POOL variable in our DI. STUDIES is one of the variables for integrated information in each line of IADSL. For this reason, we need both variables filled at the same time. With this rule, we deviate from ADAMINT convention to have STUDIES blank in lines that are not POOL=”Overall”.

COMPARISON OF USE CASE TO PHARMASUG 2018 PUBLICATION

The publication Zhong et al. [1] describes a complex data structure that was designed based on the draft ADAMINT. Their example for a complex case is the different definition of standard variables in ADSL, e.g. TRTSDT or covariates have different values depending on the pool used.
The ADaM classes IADSL and IBDS are used. As described above, IADSL contains one record per subject per pool plus one POOL = Overall record. With this set of subject level records for each pool, standard variables can have different values per USUBJID. The publication gives a good illustration of the ideas rolled out in the ADAMINT.
The data structure and methods described has restrictions when a subject was part of more than one study. As an example, treatment period variables cannot just be stacked together like a pile, with one study representing one period, but must be derived. In our data, having one subject in more than one study is rather normal than an exception. This is why we were not able to copy this structure. We adapted it to meet this need by using the STUDYID USUBJID SUBJID STUDYID combination as unique identifier for ADSL and derive the periods required as pools.

COMPARISON OF USE CASE TO PHUSE EU CONNECT 2019 PUBLICATION

The publication from Phelan [2] describes a data structure that was designed before the draft ADAMINT. After their DI, the authors got aware of the ADAMINT and investigated what they would have done differently if the ADAMINT would have been available.
The structure adhered to the standard CDISC class ADSL with one record per subject. According to their estimation, for their complex analysis the IADSL with multiple records per subject would have been appropriate.
Instead of POOL, they used COHORT variables in ADSL, which were all in the same single line of each subject. They contained information such as “Placebo” or “Treatment x”.
This is comparable to what we did in our use case for the grouping variables except STUDYID. One line per subject per study plus different POOL variables horizontally.

An interesting aspect raised in this example of DI is the use of the STUDYID variable in BDS datasets. How should it be filled in case a record for a cohort or pool is derived from more than one study? As an example, the “Treatment x” period covers not only one, but two studies. They solved the problem by using the ID of the first study the subject took part in from the perspective of the cohort. An additional variable ASTUDYID was used to hold the information from which study the actual content originated, e.g. a lab result. In case of more complex derivations, any time when data from different studies were aggregated to a new record, ASTUDYID was set to null (e.g. Time on study, averages combining measurement results to a new record).

COMPARISON OF USE CASE TO PHARMASUG 2024 PUBLICATION

The publication from Minjoe (2024) goes some further steps to a new standard for DI. After the refusal of FDA to accept the new standard with more than one line per subject in ADSL, it was clear that the information of repeated participation must be stored in a different special-purpose dataset, where more than one line per subject is allowed. The key idea was taken from SDTM standard. Domain DC (Demographics as Collected) was proposed by the CDISC SDS team in 2022. Its intention is to hold information on subjects who were included in a study more than once. SDTM.DM remained unchanged with one line per subject, but DC allowed for repeated lines per subject with the variables used as in DM. When ADaM datasets are derived from SDTM.DM and SDTM.DC, they have one record per subject in ADSL and one record per subject per participation in ADPL (Participation-Level Analysis Dataset) and BDS datasets.
As an example, in the following tables we assume that the first subject was included twice, once with SUBJID=01 and once with SUBJID=08. In ADSL, we have only the first SUBJID, in ADPL we have both. Here, we would also have all the other information for each inclusion, like treatment start, age at inclusion etc.

Table 3: Extract of ADSL and ADPL for multiple inclusion of a subject

DOMAIN	STUDYID	USUBJID	SUBJID
ADSL	STUDY1	S1-01	01
ADSL	STUDY1	S1-02	02

DOMAIN	STUDYID	USUBJID	SUBJID
ADPL	STUDY1	S1-01	01
ADPL	STUDY1	S1-01	08
ADPL	STUDY1	S1-02	02

In the publication Minjoe (2024), this idea is transferred to a DI on ADaM level. As a participation is not the same as a pool, but one participation can result in different pools, the concept must be adapted. Minjoe proposes to have ADSL with one record per subject and in addition ADIPL (Integration Pool-Level Analysis Dataset) with one record per subject per pool. The ADaM class would be ADAM OTHER. In the corresponding BDS datasets, the pool variable will be present to distinguish between measurements for different pool categories.
An interesting aspect is that the BASETYPE variable will often be equivalent to the pool variable, in case the baseline needed is different for each pool. This shows that the proposed structure is not far away from concepts we are already familiar with in CDISC ADAM structure.

Comparing our DI structure to the one of Minjoe, our IADSL is very close to the ADIPL. The main difference is that we use STUDYID instead of the POOL variable and POOL as an additional category for grouping the records according to study design. Our integrated variables which hold the same value in all lines of a subject are the equivalent to the general information in the ADSL record of the subject.

CONCLUSION

Data integrations have many variations, which results in the difficulty to find a common standard to cover them. The standard must fulfill the needs of several groups of people: programmers as well as reviewers in drug authorities may be mentioned as the most important ones. The ADAMINT experience shows that CDISC as the developer of standards is somehow dependent on the acceptance of the new ideas by the medical authorities. If they reject a submission, this is a big loss for the pharma development process. Because of this, it is always very important to discuss in advance the structure of the DI with the relevant authority.

Integrated information can be stored in three ways: new columns, new rows or a new dataset.
One element is present in all proposed DI structures, this is the usage of POOL variables in IADSL, ADSL and BDS

data sets. This concept is close to BASETYPE. The concept was introduced in ADaMIG v1.1 section 4.2, in which different BASETYPE record sets were implemented. We have the same data of a subject in different rows in case the subject's assessment data are analyzed for more than one analysis definition or analysis category. In many cases, BASETYPE will be equivalent to POOL, this makes the concept intuitive for DI.

The main steps to achieve a standard for DI are so far the ADAMINT and the publication of Minjoe.
Having USUBJID and POOL as unique identifiers is a key idea of both. It is transparent. Despite that the ADAMINT has not been published, the POOL variables will probably be used in the future for complex data analyses.
Hopefully, a new effort to develop a DI standard will be started by CDISC, considering the key requirements of the authorities and the evolving ideas from programmers.
Data integration is a huge time investment in the beginning, but in the end it saves time, if the structure is transparent, well maintainable and tailored to the desired analysis. For this reason, it is worth investing this time and checking carefully the options and experiences of others, as long as a real standard is not available.

REFERENCES

[1] ADaM Structures for Integration: A Preview, Deborah Bauer, Wayne Zhong, Kimberly Minkalis
https://pharmasug.org/proceedings/2018/DS/PharmaSUG-2018-DS03.pdf
Best Paper Award
[2] Choosing the right path to follow when integrating ADaM, PHUSE EU Connect 2019, Magalie Gallet, Laura Phelan
https://phuse.s3.eu-central-1.amazonaws.com/Archive/2019/Connect/EU/Amsterdam/PAP_SI09.pdf
[3] Is a Participation-Level ADaM Dataset a Solution for Submitting Integration Data to FDA? Sandra Minjoe
https://www.lexjansen.com/pharmasug/2024/SS/PharmaSUG-2024-SS-213.pdf
[4] ADaM Data Structures for Integration (Version 1.0 (Draft)) 2018, © 2019, CDISC
Not published.
[5] Data Integration: A theoretical Perspective, PODS 2002, Maurizio Lenzerini
https://www.diag.uniroma1.it/~lenzerin/homepagine/talks/TutorialPODS02.pdf
[6] Data Integration: https://en.wikipedia.org/wiki/Data_integration

ACKNOWLEDGMENTS

I want to acknowledge Anja Schulze, Anamaria Dinu and Jörg Sahlmann for their encouraging support during the work on this paper and the associated presentation. And the great team that worked together on this challenging use case deserves my gratefulness and appreciation. Discussing and developing the structure presented here would not have been successful without any of you.

CONTACT INFORMATION

Author Name: Renate Scheiner-Sparna
Company: Alira Health GmbH, Munich, Germany
Email: renate.scheinersparna@alirahealth.com
Website: https://alirahealth.com/