

DS11

Data Integration on CDISC ADaM Level: Multi-Purpose Structures of ADSL and BDS Datasets

Renate Scheiner-Sparna, Principal Statistical Programmer





Today's Agenda

Introduction

4

Status on Complex Data Integration Standards Development 2018-2024

9

Draft ADaM Integration Guideline

11

Use Case and Comparison to Draft ADaM Integration Guideline

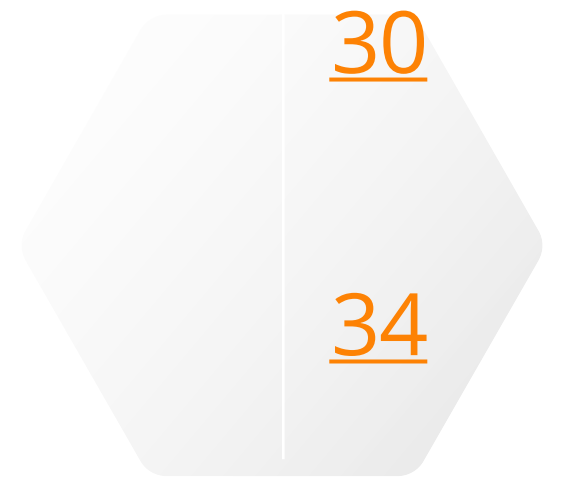
15

Proposal of a Data Integration Structure by Minjoe (2024)

30

Conclusion

34





Introduction



Introduction: Data Integration



What is a Data Integration?

- > Data integration involves combining data residing in different sources and providing users with a unified view of them
- > The data being integrated is received from a heterogeneous database system and transformed to a single coherent data store that provides synchronous data across a network of files

In the context of clinical studies:

- > Multiple studies on the same substance
- > Improve quality of analysis results by using all information available



Introduction: Data Integration



A data integration (DI) has to cover different requirements:

- > Deliver integrated summaries (ISS, ISE)
- > Support publications
- > Address requests from authorities
- > Base for company management decisions



Introduction: Data Integration



Data integration is a huge time investment

- > Single studies have differences in definitions and data capture.
- > Decisions that go beyond the standards have to be taken.
- > Data transparency and traceability back to the source data must be maintained.
- > Quality costs time.

Abbreviation **DI will be used for Data Integration in this presentation.**

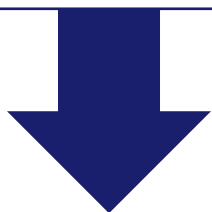


Introduction: Data Integration

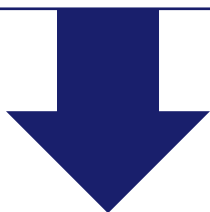


Data Integration: Simple vs. Complex

Simple	Complex
Each subject participates in one study	A subject can participate in more than one study
One specific purpose of the integration, e.g. ISE, ISS	Address different needs



ADSL and BDS will look like the ones of a single study.



More complex structures are necessary.





Complex Data Integration Standards Development 2018-2024



Complex Data Integration Standards Development 2018-2024



Since 2015	Development of ADaM Integration Guideline (ADAMINT) for ADSL
2018-2019	Development of ADaM Integration Guideline for BDS and OCCDS
2017-2019	Some publications on DI referring to ADAMINT contents
2019	End of ADAMINT review
2019	Statement of FDA: Refusal of submissions with more than one line per subject in ADSL
2019	Publication of ADAMINT cancelled
2024	Publication from Minjoe presents a new idea for DI ADaM structure, see separate section

Current status:

- > ADAMINT not available online.
- > No CDISC standard on DI available.
- > Complex DI submissions to FDA are discussed individually.





Draft ADaM Integration Guideline



Draft ADaM Integration Guideline



Throughout this presentation, we will call it **ADAMINT**.

ADSL:

New identifiers:

> POOL / POOLN

- Pooling categories
- Allow for multiple records per USUBJID in ADSL

Proposed structure in ADSL:

- > One record per USUBJID, POOL
- > One POOL=OVERALL record



Draft ADaM Integration Guideline



Other new standard variables (selected only):

> **STUDYID, STUDIES**

- One is used in each record, not both.
 - STUDYIES for POOL=OVERALL lines, STUDYID for other lines.
- Intention is traceability.

The proposed ADaM class is Multi-Purpose, the ADaM name **IADSL**.

USUBJID	POOL	POOLN	STUDYID	STUDIES
001	OVERALL	100		Study1 Study2
001	PLACEBO	1	Study1	
001	ACTIVE DRUG	2	Study2	



Draft ADaM Integration Guideline



BDS:

New identifiers:

- > **POOL / POOLN**
 - Pooling categories

Proposed structure in BDS:

- > One record per USUBJID, POOL, PARAMCD

USUBJID	POOL	POOLN	PARAMCD
001	PLACEBO	1	ECOG
001	PLACEBO	1	BPI
001	ACTIVE DRUG	2	ECOG
001	ACTIVE DRUG	2	BPI





Use Case and Comparison to Other Publications



Use Case



Use case short description:

- > Rare disease unit
- > About 10 single studies, same disease, same medicinal product
- > Phase II to IVb
- > Subjects took part in one or more study (up to 7)
- > Different data models (raw, SDTM, analysis data)

Use case intention:

- > Transparent view of all data in one structure
- > Explorative, no planned submission
- > Data should be accessible quickly for various items, e.g. safety, description of subjects, pooling of treatment phases

Use Case



DI development (2023)

- > Information available:
 - ADAMINT in review process, but not publicly available
 - We did not know that the review had been already cancelled
 - Some publications on contents of the ADAMINT

Key decisions

- > Concept of growing database: start small and flexible, add the data according to new requests
- > Complex data structure with more than one line per subject in ADSL
- > Use proposed class IADSL
- > Emphasis on study level of data → one line per USUBJID, STUDYID



Use Case Data Structure



IADSL:

- > **USUBJID** = USUBJID of study 3, if subject took part, else the one of his first study
- > **SUBJID** = SUBJID of the single study
- > **STUDYID, STUDIES**: Both are filled in each record, STUDYID is one of the identifiers
- > Treatment variables: **TRTSDT, TRTEDT** = first treatment in any study, equal in all lines per USUBJID
- > **TRT0xA, TRT0xSDT, TRT0xEDT** = actual treatment in single studies
- > **APERIODx, APERxSDT, APERxEDT** = integrated analysis periods over all studies

Unique key variables: One record per USUBJID and STUDYID

4 subjects are shown in the table.



Use Case Data Structure



IADSL (extract):

STUDYID	SUBJID	USUBJID	STUDIES	POOL	POOLC	TRTSDT	TRT01SDT	TRT01A	TRT02A	TRT03A	APERIOD1	APER1SDT	APER1EDT
Study 03	001	03-001	03	1	All verum	10FEB2022	10FEB2022	AAA		AAA	AAA	10FEB2022	03JUN2022
Study 02	002	02-002	02	2	Comparison	06JUN2018	06JUN2018	AAA		AAA	AAA	06JUN2018	02NOV2023
Study 01	002	03-005	01-03	1	All verum	09SEP2021	09SEP2021	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 03	005	03-005	01-03	1	All verum	09SEP2021	02JAN2023	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 01	003	03-024	01-02-03	1	All verum	03APR2019	03APR2019	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 02	013	03-024	01-02-03	2	Comparison	03APR2019	01DEC2020	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 03	024	03-024	01-02-03	1	All verum	03APR2019	09MAR2021	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020

4 subjects, number
from Study 03

Use Case Data Structure



IADSL (extract):

STUDYID	SUBJID	USUBJID	STUDIES	POOL	POOLC	TRTSDT	TRT01SDT	TRT01A	TRT02A	TRT03A	APERIOD1	APER1SDT	APER1EDT
Study 03	001	03-001	03	1	All verum	10FEB2022	10FEB2022	AAA		AAA	AAA	10FEB2022	03JUN2022
Study 02	002	02-002	02	2	Comparison	06JUN2018	06JUN2018	AAA		AAA	AAA	06JUN2018	02NOV2023
Study 01	002	03-005	01-03	1	All verum	09SEP2021	09SEP2021	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 03	005	03-005	01-03	1	All verum	09SEP2021	02JAN2023	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 01	003	03-024	01-02-03	1	All verum	03APR2019	03APR2019	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 02	013	03-024	01-02-03	2	Comparison	03APR2019	01DEC2020	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 03	024	03-024	01-02-03	1	All verum	03APR2019	09MAR2021	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020

4 subjects, number
from Study 03

Both filled, STUDYID
is identifier

Use Case Data Structure



IADSL (extract):

STUDYID	SUBJID	USUBJID	STUDIES	POOL	POOLC	TRTSDT	TRT01SDT	TRT01A	TRT02A	TRT03A	APERIOD1	APER1SDT	APER1EDT
Study 03	001	03-001	03	1	All verum	10FEB2022	10FEB2022	AAA		AAA	AAA	10FEB2022	03JUN2022
Study 02	002	02-002	02	2	Comparison	06JUN2018	06JUN2018	AAA		AAA	AAA	06JUN2018	02NOV2023
Study 01	002	03-005	01-03	1	All verum	09SEP2021	09SEP2021	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 03	005	03-005	01-03	1	All verum	09SEP2021	02JAN2023	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 01	003	03-024	01-02-03	1	All verum	03APR2019	03APR2019	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 02	013	03-024	01-02-03	2	Comparison	03APR2019	01DEC2020	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 03	024	03-024	01-02-03	1	All verum	03APR2019	09MAR2021	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020

First treatment in any study First treatment in single study

4 subjects, number from Study 03

Both filled, STUDYID is identifier

Use Case Data Structure



IADSL (extract):

STUDYID	SUBJID	USUBJID	STUDIES	POOL	POOLC	TRTSDT	TRT01SDT	TRT01A	TRT02A	TRT03A	APERIOD1	APER1SDT	APER1EDT
Study 03	001	03-001	03	1	All verum	10FEB2022	10FEB2022	AAA		AAA	AAA	10FEB2022	03JUN2022
Study 02	002	02-002	02	2	Comparison	06JUN2018	06JUN2018	AAA		AAA	AAA	06JUN2018	02NOV2023
Study 01	002	03-005	01-03	1	All verum	09SEP2021	09SEP2021	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 03	005	03-005	01-03	1	All verum	09SEP2021	02JAN2023	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 01	003	03-024	01-02-03	1	All verum	03APR2019	03APR2019	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 02	013	03-024	01-02-03	2	Comparison	03APR2019	01DEC2020	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 03	024	03-024	01-02-03	1	All verum	03APR2019	09MAR2021	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020

First treatment in any study First treatment in single study

4 subjects, number from Study 03

Both filled, STUDYID is identifier

Integrated analysis period 1:
Ends before NOT TREATED period from Study 02

Use Case Data Structure



IADSL (extract):

STUDYID	SUBJID	USUBJID	STUDIES	POOL	POOLC	TRTSDT	TRT01SDT	TRT01A	TRT02A	TRT03A	APERIOD1	APER1SDT	APER1EDT
Study 03	001	03-001	03	1	All verum	10FEB2022	10FEB2022	AAA		AAA	AAA	10FEB2022	03JUN2022
Study 02	002	02-002	02	2	Comparison	06JUN2018	06JUN2018	AAA		AAA	AAA	06JUN2018	02NOV2023
Study 01	002	03-005	01-03	1	All verum	09SEP2021	09SEP2021	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 03	005	03-005	01-03	1	All verum	09SEP2021	02JAN2023	AAA	AAA	AAA	AAA	09SEP2021	20AUG2023
Study 01	003	03-024	01-02-03	1	All verum	03APR2019	03APR2019	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 02	013	03-024	01-02-03	2	Comparison	03APR2019	01DEC2020	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020
Study 03	024	03-024	01-02-03	1	All verum	03APR2019	09MAR2021	AAA	NOT TREATED	AAA	AAA	03APR2019	30NOV2020

First treatment in any study First treatment in single study

4 subjects, number from Study 03

Both filled, STUDYID is identifier

APERIOD2 (not displayed): Starts with 09MAR2021 from Study 03

Integrated analysis period 1: Ends before NOT TREATED period from Study 02

Use Case Data Structure



BDS: one parameter in ADQS

Unique key variables: One record per USUBJID, STUDYID, PARAMCD, AVISIT

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



Use Case Data Structure



BDS: one parameter in ADQS

Unique key variables: One record per USUBJID, STUDYID, PARAMCD, AVISIT

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

Identifiers (POOL has a 1:1 relation to STUDYID)



Use Case Merge



Merge of BDS and ADSL

Code is dependent on the intention.

The interest is the general integrated period	The interest is in one or more of the single studies
Only one row of each subject from ADSL is needed, the integrated variables are equal in all lines.	We need all the rows of the studies of interest, e.g. all studies belonging to POOL=1 (01, 03).
Keep only integrated variables, others would cause confusion!	
<pre>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;</pre>	<pre>data IADQS_merged; merge IADQS (in=INQS) IADSL; by USUBJID STUDYID; if INQS and POOL=1 then output; run;</pre>

Use Case – Compared to ADAMINT



3 main differences:

Use case	ADAMINT
POOL variable is equivalent to STUDYID. STUDYID is used as identifier. > One record per USUBJID and STUDYID > Very intuitive	One record per USUBJID and POOL
Records for POOL= Overall are not derived > Integrated information is stored in variables. > Makes the ADSL very broad (>100 variables) > Follow naming convention.	Records for POOL= Overall are proposed as standard
Both STUDYID and STUDIES are filled.	Either STUDYID, or STUDIES is filled.



Use Case – Compared to More Publications



I planned to present more comparisons, to more publications and programming examples.

Please visit the PHUSE archive and find the paper, Time is too short to present all.





Proposal of a Data Integration Structure (Minjoe, 2024)



Proposal of a New DI Structure (Minjoe, 2024)



Main pre-assumption: One line per USUBJID in ADSL

→ Where can the integrated information be stored?

- > Core idea comes from SDTM:
 - CDISC SDS team proposed SDTM domain DC – Demographics as Collected
 - Information on subjects who were included in a study more than once
 - Structure and standard variables like domain DM, but with repeating lines for SUBJID
- > When ADaMs are derived from SDTM.DC and DM:
 - ADaM domains ADSL and ADPL – Participation-Level Analysis Dataset
 - Definitions necessary to fill the differing variables

DOMAIN	USUBJID	SUBJID	ARM	STUDYID
ADPL	001	001	PLACEBO	Study1
ADPL	001	030	ACTIVE DRUG	Study1

DOMAIN	USUBJID	SUBJID	ARM	STUDYID
ADSL	001	001	XXX	Study1

Proposal of a New DI Structure (Minjoe, 2024)



This idea is now transferred to DI.

- > Multiple participation is not the same as a pool, but can result in different pools.
- > ADSL and **AD IPL – Integration Pool-Level Analysis Dataset**
- > One record per subject per pool
- > The corresponding BDS ADaMs are the same as already presented!



Proposal of a New DI Structure (Minjoe, 2024)



Combining information from
several studies to **AD IPL**

(From the original paper,
see references)

STUDYID	USUBJID	SUBJID	TR01SDT	TRT01P
AAA	AAA-01	AAA-01	02FEB2023	A

Table 8: ADSL for Subject 1 in a Double-Blind Study

STUDYID	USUBJID	SUBJID	TR01SDT	TRT01P	TR02SDT	TRT02P
BBB	BBB-02	BBB-02	02APR2023	A	15JUN2023	B

Table 9: ADSL for Subject 2 in a Two Arm Study

STUDYID	USUBJID	SUBJID	TR01SDT	TRT01P
AAA	AAA-03	AAA-03	10MAY2023	B

Table 10: ADSL for Subject 3 in a Double-Blind Study

STUDYID	USUBJID	SUBJID	TR01SDT	TRT01P
CCC	AAA-03	CCC-05	02SEP2023	A

Table 11: ADSL for Subject 3 in an Open Label Study

Table 12 combines the 4 prior tables into a single ADIPL dataset:

STUDYID	USUBJID	SUBJID	POOLID	TR01SDT	TR01SDT
AAA	AAA-01	AAA-01	DB	02FEB2023	A
BBB	BBB-02	BBB-02	DB	02APR2023	A
BBB	BBB-02	BBB-02	OL	15JUN2023	B
AAA	AAA-03	AAA-03	DB	10MAY2023	B
CCC	AAA-03	CCC-05	OL	02SEP2023	A

Table 12: ADaM ADIPL for three subjects

DB= Double Blind, OL= Open Label



Conclusion



Conclusion



Comparing the different DI structures.

Integrated information on a subject: This is derived from more than one study, “Overall” information

	Use Case	ADAMINT	Minjoe
Key in ADSL	USUBJID STUDYID	USUBJID POOL	USUBJID In ADIPL: USUBJID POOL
Where is integrated information stored?	Additional variables in ADSL	Additional “Overall” rows in ADSL	Separate dataset ADIPL
Traceability, Transparency	> High because of study-wise information > Recommended naming convention for integrated variables	> Needs clarity for use of standard variables in pool rows > Origin of pool information not always transparent, needs high effort in specifications	> Clear separation of integrated and pool-wise information in two datasets > More experience needed, high effort in specifications
Effort / Time	Impact on methods definition high	Impact on methods definition high	Impact on methods definition high
Main advantages	> Intuitive > Flexible to different requests > Tried to apply what was expected to become the standard	> Developed further the idea of different BASETYPES > Broad acceptance on the way to become a standard > Already some examples available	> Conformance to FDA requirements > Somehow organic development from well-known CDISC concepts > Chance to be the starting point for a new standard

Conclusion



- > Data Integrations have many variations – makes it difficult to find a common standard.
- > Integrated information can be stored in 3 ways: new columns, new rows, new dataset.
- > Needs of several groups must be considered: programmers, reviewers in drug authorities.
- > Data structure must always be discussed in advance with the authorities to avoid rejection.

- > Data Integration is a huge time investment, but in the end it saves time if the structure is:
 - Transparent.
 - Well maintainable.
 - Tailored to the desired analysis.



Conclusion



If we would start today... Lessons learned.

- > We would re-derive all data from SDTM and not use previous ADaMs.
- > We would store more of the integrated information in rows, even if columns can't be avoided.
- > We would try the ADSL + ADIPL from Minjoe, because it combines the advantages of minimal structure in ADSL and Overall information.

Hopefully, the ADSL + ADIPL structure can be the starting point of a new CDISC standard. It is close to what is already used by CDISC standards, the concepts are already known. It satisfies the FDA criterion.



Conclusion



Your experience.

- > Do you find your own complex DI programming in the examples presented so far?
- > Do you have something different? We would like to hear and discuss!
- > Is there any experience with a submission to authorities?

Discussion appreciated, I am also available later by email.

Thank you very much for your kind attention!



References



[1] ADaM Structures for Integration: A Preview, Wayne Zhong, Kimberly Minkalis, Deborah Bauer

<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, Laura Phelan, Magalie Gallet

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 Guideline on Data Integration (Version 1.0 (Draft)) 2019, CDISC

Not published

[5] Data Integration definition https://en.wikipedia.org/wiki/Data_integration

[6] Data Integration: A Theoretical Perspective, PODS 2002, Maurizio Lenzerini

<https://www.diag.uniroma1.it/~lenzerin/homepage/talks/TutorialPODS02.pdf>





Q & A

Contact Me at:

|  renate.scheinersparna@alirahealth.com

Thank you, let's Keep in Touch

 alirahealth.com  @Alira Health  @AliraHealth

