Paper CD08

Thomas Clinch, Theorem Clinical Research, King of Prussia, PA, USA
Nate Freimark, Theorem Clinical Research, King of Prussia, PA, USA

ABSTRACT

When it comes to integration the SDTM Model and IG do not give very much guidance. There is much discussion about the need for ISS/ISE level SDTM datasets and how to create those versus not needing integrated SDTM. The SDS and ADaM team have a subteam that are working to address this topic, but the work is still in early development.

From an ADaM perspective there is generally a requirement for integrated analysis datasets. There are many ADaM variables that have to be modified or added to allow for integrated analyses. Whether these datasets can or should be created on a project level and just stacked together for ISS/ISE analysis or if there is a need for the creation of integrated datasets depends on the nature of the variable updates and requirements.

ADSL can present a unique set of challenges. Whether there should be one record per subject or one record per subject per study, treatment start and stop dates and groupings for analyses for variables that change over time are all issues that have to be considered when building ADSL to support integrated analyses. The question of whether to create one ADSL to support ISS and ISE or whether to have separate ADSL datasets is also a topic that has to be decided upon.

This paper will cover the various methodologies and considerations that a sponsor may look at when creating ADSL for integrated analyses. Among the topics covered will be “source data”, “one ADSL or two”, “how many records per subject” and “update existing variables or create new variables”.

INTRODUCTION

The SDTM [1] and ADaM [2] standards currently cover single study data structures in great detail. What has not yet been covered explicitly by the CDISC standards is how to combine data for an integrated analysis. Whether the target is an Integrated Summary of Safety (ISS) or an Integrated Summary of Efficacy (ISE), a great deal of planning is required to create a set of data that will support the desired analyses.

While industry and the CDISC teams are looking into some answers and guidance around integrated data, the current need for integrated data has been pointed to by the FDA review divisions [3].

In addition to ongoing considerations within the SDTM and ADaM teams, integration was the focus of one of the working groups that was identified as being needed at the FDA/PhUSE Computational Science Symposium that occurred March 19th and 20th, 2012. Working Group 3 was formed to address “Challenges of Integrating and Converting Data across Studies”. Until an industry consensus is developed, there are many decisions that must be made regarding the path that will be taken to get to the integrated analyses required.

The purpose of this paper is to cover some of the issues that arise when combining data to create an integrated subject-level analysis dataset (ADSL) from multiple studies.

The topics covered in this paper are:

1) What is the source data?
2) How many ADSLs?
3) How many records per subject?
4) Use existing variables or create new variables?

What is the Source Data?

There is much discussion about what should be the source data for integrated analysis datasets. There are at least three possibilities: study-level SDTM, integrated SDTM, or study-level ADaM. There are pros and cons for each methodology.

Study Level SDTM

Pros	Cons
• Source data for ADSL remains unchanged and matches the source data for the individual studies • No additional SDTM programming or decisions are needed	• Events that cross studies are not combined • VISIT/TPT/TESTCD and other values are not consistent • Traceability can be difficult • Recoding would need to be done at a study level, for example, bringing each study MedDRA coding up to a more recent version used in the later studies

Integrated SDTM

Pros	Cons

• Source data is updated to support ADaM values
• Define.xml can point to data within same study instead of looking back at multiple versions of the same domain across the individual studies for the ADaM value source
• Allows for capturing in SDTM data that may not be available in a single record or domain in the individual studies
  o Medical history for extension studies
  o AE end dates for roll-over studies
	• Requires an extra level of dataset creation and QC
	• Requires harmonization of SDTM variables, values, and structures
	  o VISIT/VISITNUM
	  o TPT/TPTNUM
	  o TEST/TESTCD

Study Level ADaM

Pros	Cons

• all calculated values are available
• structures are similar
• allows for availability of ADaM-derived values on an integrated level
  o study level population flags
  o disallowed medication flags
  o exposure duration/compliance values
	• Requires 100% pre-planning
	• Requires addition of variables on a study level that are not relevant on a study-level analysis
	• Assumes that all integrated ADaM will have a study-level ADaM

How Many ADSLs?

Once the source of the integrated ADSL is decided upon, the next question is how many ADSL datasets should be created. There are two levels to this question. The first is whether there should be one ADSL to support all integrated analyses (ISS and ISE) or whether there should be one ADSL for ISS and one ADSL for ISE. There are two basic differences between ISS analyses and ISE analyses. ISS analyses typically cover all of the studies where active treatment medication was administered (including Phase I trials), whereas ISE analyses generally cover studies where efficacy was assessed (typically Phase III trials). Furthermore, ISS analyses are generally fairly straightforward looking at the safety profile with very few subgroup analyses, whereas ISE analyses typically has many subgroup analyses to check for robustness of efficacy results.

Having multiple ADSL datasets does allow for more streamlined and focused ADSL since the pivotal studies that go into an ISE typically have more variables and groupings to capture baseline values for subgroup analyses. It also allows population flags to be captured in both ISS and ISE that mirror the study level populations. One of the issues that has to be dealt with is how to store these datasets. Since a single folder can only have one version of ADSL, there would either have to be naming conventions other than ADSL or there would have to be a folder to capture ISS datasets and a folder to capture ISE datasets. This would drive the need to have multiple datasets in each folder if any datasets were used for both ISS and ISE analyses and there would be a need for two versions of the define.xml if the links are based on links to datasets based on relative folders.

However, if one ADSL is created, then there will possibly be population flags that have no value for studies that did not have all populations defined (for example if PPROTFL is defined for Phase III trials and not Phase I trials) which will create compliance issues per the current rules in OpenCDISC. Furthermore, there may be different rules for the same variables for ISS analyses vs ISE analyses which would mean that there would have to be several variables to capture the same concept. For example, if ISS relative days are anchored by TRTSDT (Treatment start date) and ISE relative days are anchored by RANDDT (Randomization date) then there would possibly be the need to create both AGE and AGERAND to capture the AGE values based on each anchor date. The age summarized in the demographics summary for ISS vs ISE may differ, and AGE used in any derived variable calculations for efficacy analyses may need AGERAND instead AGE.

How many records per subject?

After determining the number of ADSL and the integrated database folder structure, the next consideration is the actual structure of the integrated ADSL. The default expectation of the ADaM model is that ADSL will contain one record per USUBJID, however the CDER Common Data Standards Issues Document, states that “Integrated summaries may contain more than one record per unique subject in the case that an individual subject was enrolled in more than one study.”[4]

Each possible structure for ADSL in an integrated database raises its own set of issues.

If the structure is one record per subject, then it would require a careful assessment of how to represent subjects who are enrolled in more than one trial. Are the treatments of each study considered different treatment periods? How are the same variables defined in each study captured in a single ADSL such as AGE and TRTSDT? If a subject received Placebo in one study and active treatment in another study, then possibly only the active treatment will be considered for analysis in an ISS. Finally, if the individual study datasets are used as source data, then

traceability will become difficult if ADSL variables are recalculated based on individual study ADSL (or other analysis dataset's) variables.

Having a structure of one record per subject per study does have the benefit of leaving the study-level ADSL "as-is" without any manipulations. However not having treatment start/stop dates that look across trials does not allow for slotting of assessments or events if integrated analyses have their own set of windowing rules that do not merely mimic the individual study rules. Furthermore, the values needed for analyses such as overall exposure and treatment compliance have no intuitive storage place.

Some of the considerations that may go into the decision of which structure to create will be determined based on the rules defined in the integrated SAP. If subjects will be looked at to determine the "patient experience" then values from multiple studies will have to be combined, presumably on one record, to facilitate analysis. If only one value per subject will be analyzed based on the subject state at the time of the first dose of active treatment medication (which may be the first or last study a subject was enrolled in), then careful consideration has to be given as to which variables should be created to capture the values needed for analyses. On the other hand, if each study is analyzed as if the subject were a 'unique' subject, then just stacking up the individual ADSL will suffice.

Use existing variables or create new variables?

This leads to the next consideration when building an integrated ADSL. Should the integrated ADSL have the same data structure as the study level ADSL or should there be an entirely new set of variables to capture the integrated values? Maintaining the same data structure does allow for the integrity of the data standard and the avoidance of duplication in cases where no changes are made for subjects who are not in multiple trials. This would presumably parallel the methodology that would be followed if integrated SDTM is created since SDTM does not allow for the creation of "new" ISS variables. However it does introduce the difficulty of tracing the "new" value back to the source variable that it was created from. On the other hand, creating a new set of ISS variables would allow for the maintenance of the original values and clear traceability is maintained. The issue is that the set of new variables would have to be defined, it would require an education curve by any user of the standard, and it would require the updating of any programs that were created based on the existing ADaM structure. Furthermore, if the integrated ADSL is based on integrated SDTM, then this would not follow the current ADaM model that any variable found in ADaM must match the value found in the corresponding SDTM. A hybrid solution might be to have multiple rows in the integrated ADSL, with rows to represent the values from the original study ADSL and added rows to capture the integrated values needed for integrated analyses.

Following are some examples of situations where these considerations arise.

AGE variable:

1) Study 1 has AGE created based on randomization date
2) Study 2 has AGE created based on enrollment date
3) Study 3 has AGE created based on first dose date
4) ISS SAP defines AGE based on first dose date

Should AGE be created and all study level value ignored? Should AGENEW be created and AGE left as-is? Should AGE be created and original values kept as AGEOLD?

Some possible considerations are:

1) If AGE is used for analysis do we care what the original study actually was?
2) Is there a desire to recreate study level results from the integrated ADSL
3) If the SDTM is integrated does this all become a moot issue since SDTM will theoretically have only AGE value (assuming that SDTM only has a rederived AGE value that matches the analysis rules)?

Baseline Values:
Study 1 has baseline calculated as the average of the windowed Screening value and the windowed Day 1 value
Study 2 has baseline defined as the nominal Day 1 value
ISS has baseline defined as the last non-missing value prior to dosing

Should xxxxBL values be calculated for the ISS ignoring the study level variables or should a new variable be created to capture the ISS ADSL value?

Treatment Variables:

Scenario #1

a. Study 1 compared High Dose vs Placebo
b. Study 2 compared High Dose vs Low Dose vs Placebo
c. ISS looks at High Dose vs Low Dose vs Placebo

Should the study level ADSL be set up to support the integrated ADSL or should the integrated ADSL have variables adjusted to account for the ISS analyses? This would be relevant to TRT01PN for example where on an integrated level the values might be
1 (High) vs 2 (Low) vs 3 (Placebo) and the values from Study 1 would either be 1 vs 3 if the values are based on the ISS choices or 1 vs 2 if the values are based on the study level analysis values.

Scenario #2

a. Study 1 compared 30 mg vs 60 mg (TRT01PN populated as 1 vs 2)
b. Study 2 compared Fed vs Fasted (TRT01PN populated as 1 vs 2)
c. Study 3 only had 20 mg (TRT01PN =1)
d. ISS included all active treatment in one “treatment group” for analysis purposes

Should the treatment group for ISS be captured in a treatment grouping variable (i.e. TR01PG1)? Should all values be captured in TRT01P? How does TRT01P get harmonized so that TRT01P vs TRT01PN is a one-to-one mapping? Or should that not be consideration when integrating different studies?

Calculated variables based on study populations.

1) Subgroup analysis is done based on baseline value (<= median value vs. > median value) where the median value will change based on combinations of studies
    a. Study 1 has median age = 35.7
    b. Study 2 has median age = 43.3
    c. ISS has median age = 41.2

2) Compliance ranges that differ from study level
    a. Study 1 summarized <60%,60%-80%,80%-100%,100%-120%,>120%
    b. Study 2 summarized <80%,80%-90%,90%-100%,100%-110%,>110%
    c. ISS summarizes <70%,70%-80%,80%-90%,90%-100%,100%-110%,110%-120%,>120%

The issue arises of whether maintaining the individual study values provides added value. This would allow clear traceability back to the original study and would allow for reproducing of individual study results using the integrated data. Two possible methods of doing this are to add variables to capture the original study values (AGE1, AGE2, AGE3) or to add rows in the dataset with one row to capture each original study observation.

CONCLUSION

In conclusion, the proper methodology for creating an integrated subject level dataset to support ISS/ISE analysis requires discussions and decisions from the standards groups (SDTM and ADaM) with input from the proper review authorities. There are many factors to consider regarding the structure and content of the required ADSL(s). And integration is complicated by the fact that factors are inter-related so any individual decision has impact in how other determinations are made.

KEYWORDS

SDTM, ADaM, ADSL, ISS, ISE, Integration, Integrated Analysis

REFERENCES

1. SDTM Model Document v1.3 and SDTMIG v3.1.3
2. ADaM Model Document v2.1 and ADaMIG v1.0
3. Guideline For The Format And Content Of The Clinical And Statistical Sections Of An Application - 21CFR314.50 (d) (5) (v) for ISE and 21CFR314.50 (d) (5) (vi) (a) for ISS
4. CDER Common Data Standards Issues Document: (http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM254113.pdf)

CONTACT INFORMATION

Author: Nate Freimark
Institution: Theorem Clinical Research
Address: 1016 West Ninth Avenue, King of Prussia, PA 19046
Telephone: 732-719-4944
E-mail: nate.freimark@theoremclinical.com

Author: Thomas Clinch
Institution: Theorem Clinical Research
Address: 1016 West Ninth Avenue, King of Prussia, PA 19046
Telephone: 859-264-9184
E-mail: thomas.clinch@theoremclinical.com