

Paper DS09

Basics of Dataset Integration

Pradeep Kumar, Quanticate International Ltd, UK

ABSTRACT

Integrations of studies are important for setting up safety and efficacy profiles of the component of interest and are referred to as ISS and ISE. Without the integration of multiple studies some larger trends in the data may not be so easily visible, regardless of whether those are good or bad, so these integrations are vital to ensuring the safety and effectiveness of any intervention.

We would like to share our experience of a recent integration using the CDISC ADaM standards along with the challenges we faced and thought processes used to resolve these to help others undergoing the same task. Traceability, a key principle of the ADaM standard, from the source datasets to our integrated datasets, was one of the primary drivers for both the issues and resolutions. Differences in the study designs, which were often complex, as were the dosing regimens, made the integrations more complex than originally hoped.

INTRODUCTION

Integrating datasets for Integrated Summary of Safety (ISS) or Integrated Summary of Efficacy (ISE) reports is common in the clinical world. In this paper we will be explaining the complexities of an integration we came across in the real world, as well as those related to implementation of CDISC ADaM guidelines for legacy studies integration and the challenges we faced while reporting. A version of implementation guide from CDISC is going to be released early next year for integration studies which is the best place to refer to before starting any work, though it may not cover all the scenarios faced during work with real data. We will be covering this whole topic from perspective of the following.

1. Purpose
2. Case study

1. PURPOSE

The purpose of doing an ISS is to establish a safety profile and an ISE is to establish effectiveness of the medication of interest. ISS and ISEs can also serve as a base for further investigations in a different therapeutic area of the same medication. When designing a study, the sample size is generally determined to allow enough power for the primary endpoint for an individual study. For ethical and cost reasons it is very rare that sample sizes are sufficiently large to allow any other efficacy and safety endpoints to be fully assessed. Integrated study data may also allow a more robust analysis of subgroups.

1.1 BENEFITS OF THE ISS:

Some of the benefits from ISS are quoted below, but they are not limited to the mentioned points.

- Identification of common related AEs
- Identification of common serious adverse events
- Identification of safety concerns that may not be apparent from a single study
- Abnormal laboratory results (e.g. liver events)
- Allowing a more robust assessment of safety in subgroups if sample sizes are sufficiently large

1.2 BENEFITS OF THE ISE:

Some of the benefits from ISE are quoted below, but they are not limited to the mentioned points.

- Enable the assessment of efficacy in subgroups (e.g. age groups, sex, race, countries)
- Enable the assessment of the efficacy of secondary endpoints, which may have been underpowered in individual studies

- Allow detailed comparison of results versus a comparator drug
- Allow assessment of the efficacy of repeat doses
- Enable the dose-response profile to be more accurately estimated
- Allow any changes that may have occurred in efficacy over the lifetime of the drug development to be investigated
- Help to address why results for a study may be inconsistent with the results of other studies

2. CASE STUDY

In the study we reported, there were 24 early phase (Phase 1 and Phase 2) studies which were reported in legacy standards containing 800+ subjects combining all studies. All subjects were treated with different medications along with the medication of interest at different dosing levels. Studies ranged from one to six periods. The summaries were based on whether a subject received a single or multiple dose of the medication along with or without concomitant medications and a placebo. The other study we reported contains 500+ subjects, which combines 2 phase 3 studies. The summaries were based on treatment groups under which subjects were randomized; both source studies were reported to CDISC ADaM standards. Some subjects were excluded if they never received the medication of interest from the summaries, but not excluded from SDTM/ADaM data. All the legacy studies data was converted to CDISC SDTM datasets, so the only focus was on the CDISC ADaM datasets. The below sections will explain in detail on the planning and handling of such vast studies integration whilst conforming to CDISC ADaM guidelines.

2.1 STUDY DESIGN

Understanding the study designs of all the studies to be integrated was the key for developing the specifications for the datasets. Equally important was identification of at which stage the medication of interest was administered. Planning of any such integrations beforehand is helpful. We need to understand the complexity of the submission and keep the details in hand such as the number of studies to be integrated, exclusions were needed to be applied in the datasets, and end points of the submission. Parallel and crossover studies can be integrated, but care must be taken when grouping doses and period definition clarified for this. A list of variables was created conforming with the CDISC ADaM naming convention which are called with the prefix of "IS". Example variables are ISPERIOD, ISSTD, ISENDT, ISS01A, ISS02A etc., The below table explains about how this information was included/excluded for summaries without losing any of the detail.

STUDYID	USUBJID	TRT01A	TRT02A	ISS01A	ISS02A
111	111.001	X	A	A	
111	111.002	A	X	A	
111	111.003	A	B	A	
111	111.004	A	AB	A	AB
112	112.001	PLACEBO		PLACEBO	

In the above Treatment A is the medication of interest, so the TRTxxA variables represent the information to match the source studies and ISxxA variables are used for integration. 111 is a cross over study with TRTA, TRTB, TRTX and TRTAB. For IS period or treatment we use TRTA reference only, ignoring treatment X and treatment B.

Whereas in some studies some subjects undergone a pre-treatment for 1 or 2 weeks with a different medication, in such cases all the data from those weeks is also excluded from summaries. If a subject withdrew in that period, then they were not counted.

Once the study designs and exclusions are understood then the work can be planned out carefully on the ISS/ISE datasets. Generally, all this detail is should be covered in the ISS/ISE SAP.

2.2 DICTIONARY VERSIONS

In the integration it is often found that some studies were completed using different versions of MedDRA and WHO drug dictionaries. Following FDA requirements, all coded terms need to be recoded to the latest dictionary (latest available dictionary version prior to submission) whilst maintaining the original coding for traceability back to the original study terms. It is challenging when integrating 20+ studies as all terms need to be checked and compared back to the individual studies.

2.3 POPULATION DEFINITIONS

Populations created for the individual study reporting may not support the integrated summaries or analyses that are required for an ISS or ISE and as such the safety population definition was re-defined in relation to the administration of medication of interest. This may or may not provide the same number of safety subjects to that of individual studies. The original study population flags were retained in the combined ADaM data sets in order to maintain traceability back to the original studies, so any new population flags were created using an IS prefix.

2.4 BASELINE DEFINITION

Baseline definition may be different across the studies, so for the needs of the integration, the baseline was re-defined to consider the last assessment prior to the administration of medication of interest or placebo. This may be different in some studies, where some subjects were administered with other medication and considered baseline before receiving such medication. Again, for there to be traceability back to the original studies the original baseline was kept. A new derived baseline was set up based on the requirements with the usage of Basetyp. Basetyp variable in ADaM to identify the original baseline from the source study and the updated baseline rows for integration.

2.5 PLANNING OF ANALYSIS DATASETS

A thorough review of the required TFL shells is needed to understand how many analysis datasets are required for the ISS/ISE submission. In our case study these were ADSL, ADLB, ADVS, ADCM, ADAE and ADEFF.

We also need to understand if there are any additional customized MedDRA queries (CQNAM) variables needed for the submission, if yes then we need to keep the original study queries along with the additional queries to ADAE dataset. There is a limitation on adding occurrence flag variables based on preferred terms in ADAE guidelines and in our study, we had first occurrence customized query flags, so we had to adjust the names as AOCCQxFL etc.

All covariates or subgroups required for the reporting should be described clearly in the ISS/ISE plan; these are then retained in ADSL along with any other baseline characteristics.

2.6 TRACEABILITY

Traceability is one of the key basics of integration of multiple studies. All analysis datasets and the derived variables must be easily traceable. Keep the original study content as much as possible so the regulatory authorities can trace back to the original individual studies. There should be the possibility to match back on the individual studies summaries to make sure that there are no changes to the content of the data in the process of reporting ISS/ISE. Documentation of all derivations in the metadata and the well versed ADRG support the traceability. If any records were excluded from the source study, this must be documented in ADRG with proper explanation. An additional document was also created where discrepancies between source study and the integrated data as a part of our quality checks process was documented to make sure no subject was excluded, and all subjects were reported as per the expectations.

2.7 CHALLENGES

Apart from the challenges mentioned above in each section more challenges were encountered when integrating laboratory results into the ADLB domain. The dataset contained about 600,000 records, which was challenging but made easier by splitting the dataset based on the laboratory categories and produced datasets such as ADLBCHM, ADLBHAEM etc. Basetyp records were also retained for each group of subjects based on the baseline definition. LOCF concept was implemented on some parameters where in such cases there are chances of getting a smaller number of LOCF when adding rows to the dataset. Checks were also made with regards to the ANLxxFL flag to make sure the original study LOCF records were not affected in the process.

Baseline was defined based on the overall daily dosage taken which was classified in this submission as the safety cohort, and not to be confused with the safety population. Additional rows were created for this subgroup in the dataset along with the original study records.

2.8 PINNACLE21 ERRORS:

There were some validator issues when the data sets were run through Pinnacle 21. Most of these were because of the checks which are not included in P21 for integrated studies. Some of the examples related to naming conventions used in the ADAE dataset. Names used for occurrence flags were not in line with the P21 checks but were needed for the summaries. In the ADRG all these discrepancies were documented with an explanation. Some of the issues were related to the naming convention of the treatment groups. The treatments were grouped irrespective of the

period, but the validator checks the treatment group naming convention based on period. It is not necessary that the validator issues be fixed when working on integrations, rather there is a need to investigate the rationale and stick with the guidelines as much as possible.

CONCLUSION

Planning upfront and understanding the studies to be integrated is the key success of any ISS/ISE submission. At the time of our study integrations there were no CDISC ADaM guidelines for integrated dataset structures, so we had to stick with the basic principles of CDISC ADaM guidelines such as one proc away and traceability etc. Always take the approach that will best allow us to stick with the basics of ADaM, while still being logically correct.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Pradeep Kumar

Company: Quanticate International Ltd

Email: Pradeep.kumar@quanticate.com