Paper TA05

Programming Approaches for Integrated Summary of Efficacy (ISE) Development in Oncology

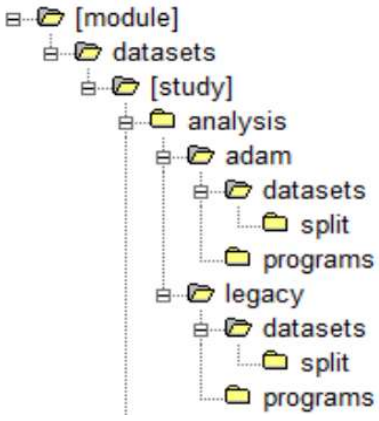
Narendra Babu Kolisetty, Merck & Co., Inc., North Wales, PA, USA
Saigovind Chenna, Merck & Co., Inc., North Wales, PA, USA

ABSTRACT

In clinical trials, the Electronic Common Technical Document (eCTD) is the standard format for submitting regulatory applications including new drug applications (NDA) and biologics license applications (BLA) to the FDA. An integrated summary of efficacy (ISE) is one of the critical components of an eCTD, providing a comprehensive overview of an investigational product's efficacy across multiple studies. However, integrating efficacy data from late-phase clinical trials into a high-quality ISE data package can be challenging due to differences in study design, data collection, and mapping procedures. Further challenges include differences in SDTM and ADaM implementation guides, inconsistent analysis result derivation, handling of external data, dealing with missing values, Pinnacle21 issues, and managing value truncations. This paper aims to provide a comprehensive summary of effective programming approaches that will enable programmers to overcome these challenges and develop high-quality ISE data packages.

INTRODUCTION

Integrated Summary of Efficacy (ISE) is a crucial element of an Electronic Common Technical Document (eCTD) submitted to the Food and Drug Administration (FDA) for new drug (NDA) or biologics license (BLA) applications. The eCTD's Module 5, section 5.3.5.3, titled "Reports of analyses from multiple studies," should include integrated data from multiple studies. Figure 1 and Table 1 summarizes the file folder structure, and their associated descriptions respectively for ISE datasets.



Folder Name	Folder Level	Description/Contents
[module]	1	Refers to the eCTD module in which study data are being submitted. Name this folder m4 for nonclinical data and m5 for clinical data. Do not place files at this level.
datasets	2	Resides within the module folder as the top-level folder for study data (nonclinical or clinical) being submitted for the specified module (m4 or m5). Do not place files at this level.
[study]	3	Name this folder with the study identifier or analysis type performed (e.g., study123, iss, ise). Do not place files at this level.
analysis	4	Contains folders for analysis datasets and software programs; arrange in designated level 6 subfolders. Do not place files at this level.
adam	5	Contains subfolders for ADaM datasets and corresponding software programs. Do not place files at this level.
datasets	6	Place ADaM datasets in this subfolder.
split	7	Place any split ADaM datasets in this subfolder.
programs	6	Place software programs for ADaM datasets, tables and figures in this subfolder.
legacy	5	Contains legacy formatted analysis datasets and corresponding software programs. Do not place files at this level.
datasets	6	Place legacy analysis datasets in this subfolder. In m4 place tumor.xpt and its associated define.pdf in this folder.
split	7	Place split legacy analysis datasets in this subfolder.
programs	6	Place software programs for legacy analysis datasets, tables and figures in this subfolder.

Figure 1

Table 1

The Integrated Summary of Effectiveness (ISE) provides a comprehensive evaluation of a study drug's benefit-risk profile, extending beyond the findings of individual clinical trials. It consolidates the drug's efficacy across various population subgroups, considering demographics, intrinsic and extrinsic factors, dosing regimens, and dose-response relationships. The ISE enhances the accuracy of results due to its larger sample size and the application of robust statistical methods for efficacy data analysis. Furthermore, the ISE contributes to the dosage and administration section of the drug labeling.

Drug efficacy data from multiple clinical trials can be integrated through several methods. These include using SDTM data alone, ADaM data alone, or a combination of both SDTM and ADaM data. Developing integrated ADaM datasets directly from the SDTM data of input studies, as illustrated in Figure 2, proves beneficial when study-level ADaM datasets are unavailable. Alternatively, one can build integrated SDTM datasets first and then derive integrated analysis datasets, as shown in Figure 3. This method aids in identifying issues at the source data level but demands significant effort, time, and resources. ADaM datasets can be integrated directly using the analysis datasets from individual clinical trials, as depicted in Figure 4. This method is less tedious and more efficient compared to others However, if some or all the required ISE analysis variables are not present in the input ADaM datasets, using both SDTM and ADaM data from individual studies would be more effective, as demonstrated in Figure 5.

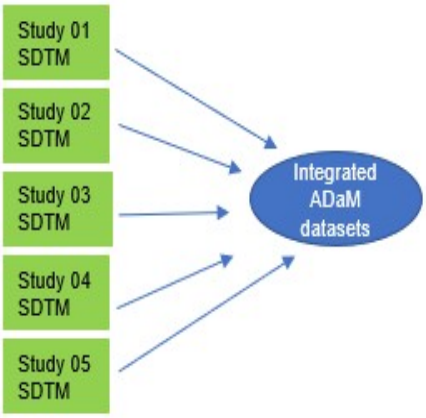


Figure 2

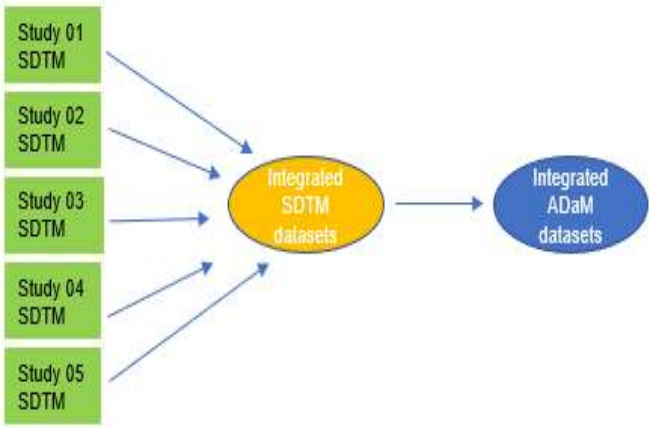


Figure 3

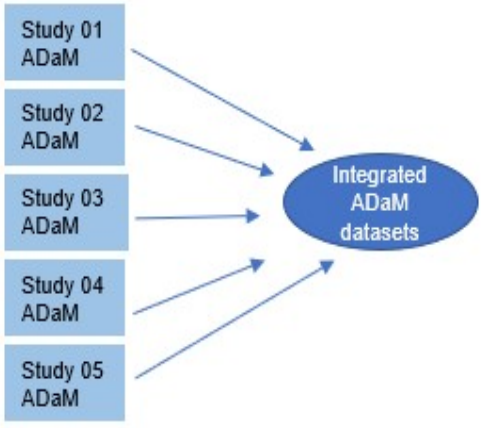


Figure 4

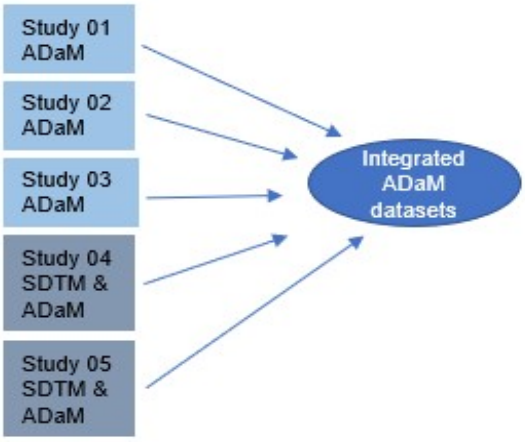


Figure 5

This paper aims to provide a comprehensive overview of efficient programming strategies for integrating efficacy data at the ADaM level. These strategies will help programmers overcome the following challenges and to develop a high-quality ISE data package.

ISE PROGRAMMING APPROACHES

TUMOR LESION DATA

Solid tumor lesion measurement data from participants in late-stage clinical trials typically present in one of efficacy datasets like Analysis of Tumor Lesions Dataset (ADTL) dataset formatted in a Basic Data Structure (BDS) at ADaM level. Figure 6 was constructed using the 'Minimum Sum of Target Lesions in Longest Diameter' parameter in ADTL dataset. As evident, analysis result value (AVAL), baseline value (BASE), change from baseline value (CHG), and percent change from baseline value (PCHG) variables were not derived for Minimum Sum of Target Lesions in Longest Diameter parameter in some of input studies for ISE. Suppose any analysis on target or non-target or new tumor lesion data is planned during integration but the needed variables like AVAL, BASE, CHG, PCHG etc. are not present in input studies, then these variables can be derived in ISE datasets utilizing SDTM data from the input studies.

In addition, if either target or non-target tumor lesion measurement data is missing at any of visits in the source data from input studies, then it's recommended to check whether a query was sent to the study site and obtained the reason for missing data. Any identified data-related Issues that cannot be fixed and there is no impact to analysis must be clearly documented in the Analysis Data Reviewer's Guide (ADRG).

STUDYID	PARAM	PARAMCD	AVAL	BASE	CHG	PCHG
Study 01	Minimum of Sum of Target/Index Lesions in Longest Diameter	MSTLDIAM	30	55	-25	-45.4545
Study 02	Minimum of Sum of Target Lesions in Longest Diameter	MSTLDIAM	100	105	-5	-4.7619
Study 03			.	.	.	.
Study 04	Minimum of Sum of Target Lesions in Longest Diameter (Radiology)	RADMSMTL	55	125	-75	-60
Study 05			.	.	.	.
Study 06	Minimum of Sum of Target Lesions in Diameter Based on Central Radiology Assessment (m...	MSTLDIRC	43	30	13	43.3334
Study 07	Minimum of Sum of Target Lesions in Longest Diameter (Radiology)	RADMSMTL	96.5	67	29.5	44.0299
Study 08	Minimum of Sum of Target Lesions in Longest Diameter Per RAD	MRADTLM	67	89	-22	24.7191
Study 09	Minimum of Sum of Target Lesions in Longest Diameter per BICR	IRCMSTL	30	34	-4	11.7647
Study 10	Minimum of Sum of Target Lesions (BICR per RECIST 1.1)	MSTLMIRC	50	57	-7	12.2807

Figure 6

TUMOR RESPONSE DATA

In late-phase oncology clinical trials, one of the primary or secondary efficacy endpoints is the participant's overall response to the treatment administered by the study drug. The response data from participants is typically presented in one of the efficacy analysis datasets like ADEFF, ADRS, ADORR, and ADIRC, formatted in a Basic Data Structure (BDS). However, the dataset's names and the values of the parameter and parameter code variables may differ between studies. Figure 7 below was compiled from multiple clinical trials to determine the best overall response in a solid tumor based on an independent review committee's assessment per RECIST 1.1. Notably, similar response assessment data was presented with varying PARAM, PARAMCD, and PARAMN values across studies. These differences necessitate additional programming work during integration of efficacy data, such as renaming PARAM, PARAMCD, and PARAMN values, redefining variable attributes, and adjusting formats etc. Programmers need to verify the code list values for character and their corresponding numeric variables. AVALC variable signifies the analysis result in a character format, while the AVAL variable supplies the corresponding numeric value. As depicted in figure 7, AVAL variable contains different numeric values for an identical character value in the AVALC variable. This discrepancy can be resolved by re-deriving AVAL values while developing integrated datasets.

STUDYID	PARAM	PARAMCD	PARAMN	AVALC	AVAL
Study 01	Best Overall Response per RECIST 1.1 (IRC Radiology Only)	BORRT	111	PR	.
Study 02	Best overall response with confirmation (IRC)	BORCFIRC	22	PR	3
Study 03	Best Overall Response (IRC) with Confirmation	BORCFIRC	22	PR	3
Study 04	Best Overall Response per RECIST1.1 (Radiology)	RADBOR	22	PR	2
Study 05	Best overall response with confirmation (Central Radiology per RECIST 1.1)	BORCFIRC	22	PR	2
Study 06	Best Overall Response with Confirmation Based on RECIST 1.1 per Central Radiology Assessment	BORCIRC	202	PR	3
Study 07	Best Overall Response per RECIST1.1 (Radiology)	RADBOR	22	PR	2
Study 08	Best overall response confirmed - RECIST 1.1, Imaging Vendor	BORRAD	21	PR	.
Study 09	Best overall response with confirmation (BICR per RECIST 1.1)	BORCFIRC	22	PR	3

Figure 7

LEGACY DATA

For integrated analysis, certain instances may necessitate using the data beyond SDTM and/or ADaM datasets from input studies. This data includes biomarker, tumor mapping, or genotype/phenotype data etc. These data files may not be received via a clinical data repository but as an external data file such as Excel, CSV, or .dat. These external data files that contain all ISE participants data will be used as an input for ISE datasets development and analysis. This includes deriving new variables such as new population flags, grouping various tumor types, classifying biomarker data, generating TLFs based on tumor type, genotype, and biomarker status etc. External data files that were received for ISE should be converted to a submission-specified format, such as xpt. Alternatively, they can be appended to the ISE ADRG, depending on the data type and scope of analysis.

Figure 8 illustrates the derivation of new variables in ISE ADSL dataset based on external data files.

Variable Name	Variable Label	Type	Length	Origin	Define Derivation
BLBMRKR	Baseline Biomarker	integer	8	Assigned	Merge the external data file "biomarker.xlsx" to the ADSL dataset by subject. Assign the values from result column by subject.
BLBMRKS	Baseline Biomarker Status	Char	40	Derived	BLBMRKS="Present", if the result result value is >10. BLBMRKS="Not-present", if the result value is not missing and <=10 BLBMRKS value is null if no result value
BLBMRKSN	Baseline Biomarker Status (N)	integer	8	Assigned	Mapped per codelist values
TMRGR	Tumor Type Grouping	Char	100	Derived	Merge the external data file "tumortype.dat" to the ADSL dataset by tumor type. Count participants by each tumor type. TMRGR="Other", if the tumor type count is <=20 Otherwise, TMRGR=TUMOR from tumortype.dat file.
TMRGRN	Tumor Type Grouping (N)	integer	8	Assigned	Sort unique tumor type values alphabetically first, and then assign numeric values from 1 to N, by interval 1. TMRGRN=999, if TMRGR= "Other".

Figure 8

PARTICIPANTS SELECTION CRITERIA

If the integration is intended to include only a subset of the total participants from the input studies, then participants selection criteria for ISE must be clearly documented in the ISE ADRG. For example: Integration is performed only on the data from participants that were received the treatment as a first-line therapy. In addition, creating a new population flag variable during the development of ISE datasets is advisable to identify list of participants, and to analyze the efficacy data. Sometimes sub-group efficacy data analysis might be conducted to compare drug's efficacy results. Example by dosing regimen, by tumor type.

ADaM and SDTM IMPLEMENTATION GUIDE DIFFERENCES

Suppose datasets are integrated to utilize ADaM datasets directly from input studies. In that case, the adherence to ADaM and SDTM IGs at the individual study level should be confirmed. If these differ from the planned approach for ISE, harmonizing variables for ISE analysis becomes essential. All required modifications, including variable name, label, formats, control terminology, and derivations, should be implemented in accordance with the ADaM IG designated for the ISE as shown in Figures 9, 10, 11, and 12.

Figure 9 illustrates the harmonization of control terminology for the RACE variable in the ISE ADSL dataset.

STUDYID	RACE	RACE_NEW
Study 01	MULTIRACIAL	MULTIPLE
Study 02	MULTIPLE	MULTIPLE
Study 03	Multi-racial	MULTIPLE
Study 04	Multi-Racial	MULTIPLE

Figure 9

Figure 10 illustrates the SRCDOM variable label and derivation as per ADaM IG version 1.0.

ADaM IG	Variable Name	Variable Label	Define Derivation
1.0	SRCDOM	Source Domain	2-letter SDTM dataset name that relates to the analysis value

Figure 10

Figure 11 illustrates the changes to SRCDOM variable label and derivation compared to ADaM IG version 1.0.

ADaM IG	Variable Name	Variable Label	Define Derivation
1.1	SRCDOM	Source Data	ADaM or SDTM dataset name that relates to the analysis value

Figure 11

Figure 12 illustrates the derivation of an actual arm variable in the ISE ADSL dataset due to difference in SDTM IG among input studies. In general, actual arm variable in ADSL dataset is a predecessor variable from SDTM DM domain. Actual arm information is mapped to ACTLARM variable per SDTM IG 311, and to ACTARM variable as per SDTM IG 312 or later versions. If there is a difference in SDTM IG between input studies, then ACTARM variable in ISE ADSL dataset needs to be re-derived as shown in Figure 12.

Variable Name	Variable Label	Type	Length	Origin	Define Derivation
ACTARM	Description of Actual Arm	Char	40	Derived	For Study 01, 03, 04: ADSL.ACTARM. For Study 02, 05, 06: ADSL.ACTLARM

Figure 12

NULL VALUES

A cautious approach is essential during the development of ISE datasets to handle variables containing null values. These null values may be reported in various ways in the input studies. Figure 13 below demonstrates the presence of null values in both character and numeric variables from different input studies for ISE. The character variable VAR1 contains values such as "Missing", "Null", "Not Reported", or a blank value. The numeric variable VAR1N has values like 99, 999, 9999, and 99999 to represent numeric null values. It is recommended to consolidate these differences into one by consistently assigning a unique character and corresponding numeric value across all ISE datasets, as depicted under VAR1_NEW and VAR1N_NEW respectively in Figure 13.

STUDYID	VAR1	VAR1N	VAR1_NEW	VAR1N_NEW
Study 01	null	999	Null	999
Study 02	not reported	999	Null	999
Study 03	MISSING	9999	Null	999
Study 04		.	Null	999
Study 05	NULL	99	Null	999
Study 06	NOT REPORTED	99999	Null	999

Figure 13

TRACEABILITY VARIABLES

Metadata traceability variables SRCDOM, SRCVAR, and data point traceability variable SRCSEQ offer data point traceability from SDTM to ADaM datasets. If the value of the AVAL/AVALC/ADT variable is directly carried forward from either an SDTM or an ADaM dataset, it is advisable to include these variables for traceability. However, these traceability variables may only be present if an individual study analysis adhered to ADaM standards.

VALUE TRUNCATION

Assigning a length value to a character variable at the beginning of a data step is essential to avoid truncation of the variable's value when setting up multiple datasets with common variables. Below figures 14 and 15, provide an overview of CNTRYNAM variable values in the output dataset with or without a length statement.

```
data adsl (keep=studyid cntrynam);
 set adsl1 adsl2 adsl3 adsl4 adsl5 adsl6 adsl7;
run;
```

```
data all (keep=studyid cntrynam);
  length cntrynam $20.;
     set adsl1 adsl2 adsl3 adsl4
         adsl5 adsl6 adsl7;
run;
```

STUDYID	CNTRYNAM
Study01	Brazil
Study02	Canada
Study03	Chile
Study04	Czech
Study05	Italy
Study06	Japan
Study07	Korea,

Figure 14

STUDYID	CNTRYNAM
Study 01	Brazil
Study 02	Canada
Study 03	Chile
Study 04	Czech Republic
Study 05	Italy
Study 06	Japan
Study 07	Korea, Republic of

Figure 15

However, if the underlying informats and formats are not cleared, the assigned length value may have a limited impact in preventing value truncation. Please refer to below figures 16 and 17 for the difference in USUBJID variable values in the output dataset with or without clearing formats.

```
data adsl;
 length studyid $10 usubjid $40 ;
set adsl01 adsl02 adsl03 adsl04 adsl05
adsl06 adsl07;
run;
```

```
data adsl;
 length studyid $10 usubjid $40 ;
  set adsl01 adsl02 adsl03 adsl04
      adsl05 adsl06 adsl07;
   format _all_;
    informat _all_ ;
run;
```

STUDYID	USUBJID
Study 01	1111-001-100001
Study 02	1111-001-100002
Study 03	1111-001-100003
Study 04	1111-001-100004
Study 05	1111-001-100005
Study 06	1111-001-100006
Study 07	1111-001-100007

Figure 16

STUDYID	USUBJID
Study 01	1111-001-1000011
Study 02	1111-001-100002
Study 03	1111-001-100003
Study 04	1111-001-1000042
Study 05	1111-001-100005
Study 06	1111-001-100006
Study 07	1111-001-100007

Figure 17

eSUB CHECKS

The ISE ADaM eSUB package comprises SAS datasets in .xpt format, define.xml, Analysis Data Reviewer's Guide (ADRG), and program files (.txt). Conducting ADaM compliance checks on ISE ADaM datasets and define file using the Pinnacle21 tool is advisable. There should not be any technical rejection criteria. Define file's derivation column should only include variable's derivation algorithm. Classify the variable's origin as either predecessor, assigned, or derived. Consistency should be maintained between the define and datasets. Any identified data-related Issues that cannot be fixed and have no impact to the analysis must be clearly documented in the ADRG.

CONCLUSION

Integrating efficacy data from multiple late-phase clinical trials into a high-quality ISE data package can be challenging for programmers and statisticians. These challenges can be mitigated by following an effective programming strategy. This paper summarized some efficient programming approaches for integrating efficacy data at the ADaM level. These strategies will help programmers to develop a high-quality ISE data package.

REFERENCES

FDA Study Data Technical Conformance Guide https://www.fda.gov/media/153632/download
FDA Study Data Standards Catalog https://www.fda.gov/industry/fda-data-standards-advisory-board/study-data-standards-resources

ACKNOWLEDGMENTS

The authors would like to thank Anand Jayaraman, Himanshu Patel, Madhusudhan Reddy Papasani, Mary Varughese, and Amy Gillespie for reviewing the paper and providing great suggestions.

CONTACT INFORMATION

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

Author Name: Narendra Babu Kolisetty
Company: Merck & Co., Inc.
Address: 351 N Sumneytown Pike, North Wales, PA 19454
Work Phone: +1 (267) 305 0403
Email: narendra.kolisetty@merck.com

Author Name: Saigovind Chenna
Company: Merck & Co., Inc.
Address: 351 N Sumneytown Pike, North Wales, PA 19454
Work Phone: +1 (267) 305 2162
Email: saigovind.chenna@merck.com