SESUG Paper DCMM-50-2025

Enhance Flexibility and Variability by Utilizing an Intermediate Dataset when Developing the ADaM Efficacy Dataset in Vaccine Studies

Edward Boykis, Ying Zhang, Ziqiang Chen, Peng Wan, Merck & Co., Inc., Rahway, NJ, USA

ABSTRACT

The increasing demand for vaccine studies necessitates rigorous efficacy analysis to ensure their efficiency and safety. This paper introduces a method that emphasizes flexibility and ease in generating diverse efficacy endpoints, which is essential for estimation and hypothesis testing for efficacy endpoints. By enabling the straightforward incorporation of various source datasets (e.g. symptoms, signs, testing results, and severity measures), this approach allows researchers to define efficacy endpoints comprehensively. The proposed methodology aims to streamline the analysis process, facilitating timely and broad evaluations that can adapt to the dynamic landscape of vaccine development and public health needs.

INTRODUCTION

In vaccine studies, the Efficacy Analysis Dataset (ADEFF) within the Analysis Data Model (ADaM) framework is used to store derived efficacy endpoints. This dataset helps organize and prepare data for statistical analysis of a vaccine's efficacy, such as measuring the proportion of participants who achieve a certain level of protection or who experience a specific outcome. These endpoints are calculated or transformed from source data and represent the key measures of vaccine efficacy.

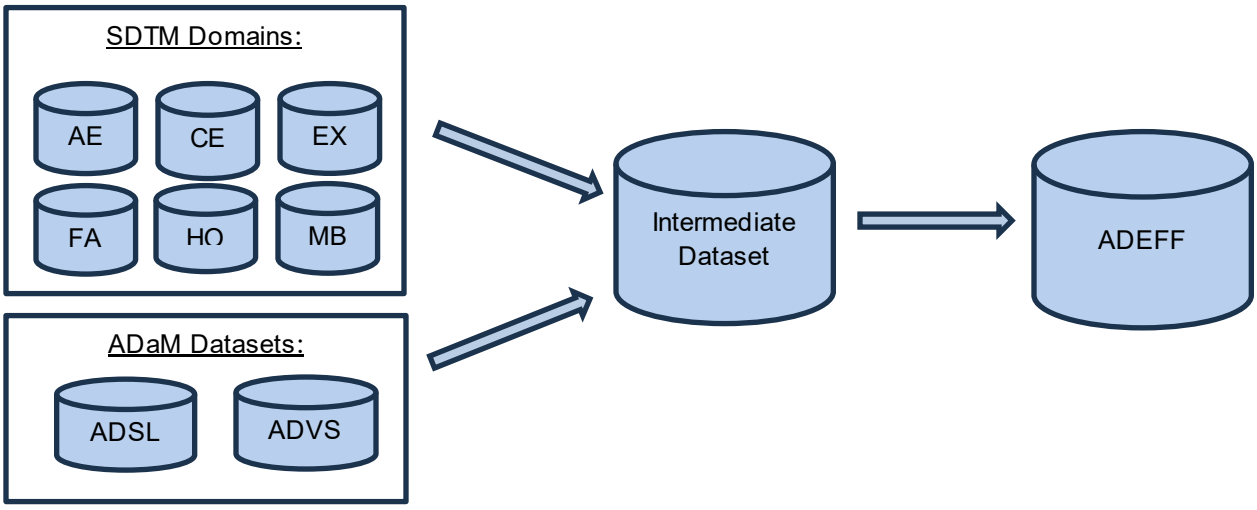
The proposed method for developing the ADEFF dataset involves two primary steps:

1. Gathering all essential elements of efficacy endpoints (such as symptoms, signs, testing results, severity measures, dates, and auxiliary flags) into a single intermediate efficacy dataset.
2. Establishing endpoint records (case/non-case) according to the distinct characteristics of each endpoint.

This approach facilitates accelerated creation or modification of efficacy endpoints, thereby optimizing time efficiency and enhancing the scalability of the final dataset.

Figure 1 illustrates the collection of efficacy data in an intermediate dataset, followed by the creation of ADEFF based on this dataset.

Figure 1: Efficacy data flow diagram



INTERMEDIATE DATASET DESIGN AND SOURCE DATASETS

To capture all relevant information in intermediate dataset, the following source datasets were used: several Study Data Tabulation Model (SDTM) domains and two ADaM datasets.

- Adverse Events (AE) domain includes clinical data describing "any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with this treatment".

- Clinical Events (CE) domain used to capture clinical events of interest that would not be classified as adverse events. The data may include signs and symptoms of the disease to be studied, or events that do not constitute adverse events in themselves, though they might lead to the identification of an adverse event.

- Findings About (FA) domain used to capture the findings about an event or intervention that cannot be represented within an event or intervention record or as a supplemental qualifier.

- Exposure (EX) domain records the details of a subject's exposure to protocol-specified study treatment. Study treatment may be any intervention that is prospectively defined as a test material within a study and is typically but not always supplied to the subject.

- Healthcare Encounters (HO) domain includes inpatient and outpatient healthcare events (e.g., hospitalizations, nursing home stay, rehabilitation facility stays, ambulatory surgery).

- Microbiology Specimen (MB) domain is designed to store microbiology findings that include organisms found, gram stain results, organism growth status, test results from assays (e.g. Real Time-Polymerase Chain Reaction (RT-PCR)).

- Subject-Level Analysis Dataset (ADSL) contains variables such as subject-level population flags, planned and actual treatment variables, demographic information, randomization factors, subgrouping variables, stratification factors, geographic information (e.g. altitude, hemisphere) and important dates.

- Vital Sign Analysis Dataset (ADVS) contains one or more records per subject, per analysis parameter, per analysis timepoint. Age at each assessment was derived in ADVS and used for efficacy endpoint determination.

With ADaM IG, the use of intermediate analysis datasets has been recommended to improve the understanding of data flow when complex data transformations are needed in support of statistical analysis. From ADaM IG v1.1:

"Very complex derivations may require the creation of intermediate analysis datasets. In these situations, traceability may be accomplished by submitting those intermediate analysis datasets along with their associated metadata. Traceability would then involve several steps. The analysis results would be linked by appropriate metadata to the data which supports the analytical procedure, those data would be linked to the intermediate analysis data, and the intermediate data would in turn be linked to the source SDTM data"

The intermediate dataset contains integrated elements needed to define efficacy endpoints such as clinical episode information (CE domain), symptoms, signs, severity indicators, clinical facility locations (FA domain), hospitalization episode information (HO domain), oxygen saturation (hypoxemia), respiratory rate (tachypnea), body temperature (fever). The dataset is organized in way one record per subject, per clinical episode (start/end dates) and per clinical facility location. Keeping all components in one dataset provides necessary traceability between source datasets (SDTM domains, ADaM datasets) and final efficacy data. With the use of the intermediate dataset, it is straightforward to understand the criteria meeting an endpoint and convenient for users to quickly identify information.

The intermediate dataset collects information from different SDTM domains and ADaM datasets as stated above by combining related records:

1. Identify signs or symptoms at each assessment for each clinical episode

• Merge CE and FA domains by CESPID=FASPID and identify symptoms such as cough, stuffy nose, and trouble feeding.

• Merge with ADVS by FADTC=VSDTC to identify vital sign related symptoms, such as tachypnea (using respiratory rate), hypoxemia (using oxygen saturation), and fever (temperatures).

2. For inpatient assessments, combine with AE and HO domains to identify the hospital admission date for hospitalization.

3. For each assessment, define potential efficacy cases based on the required combinations of symptoms for multiple endpoints.

Each record in the intermediate dataset also includes multiple flags that signify whether the specific timepoint meets the necessary symptom criteria and falls within the appropriate date ranges for the efficacy endpoints. The

Figure 2 illustrates a snapshot of intermediate dataset regarding symptom information and corresponding episode dates. This information helps to simplify the programming logic for the endpoints in the final ADEFF dataset. The Figure 3 illustrates a snapshot of endpoint flags in the intermediate dataset. At present, the dataset is not directly utilized to generate any analyses and tables/listings/figures (TLF) outputs.

Figure 2: Intermediate dataset snapshot with dates and symptoms

USUBJID	CESTDT	CEENDT	...	HOENDT	ALTITGR	SYMP01	SYMP02	...	SYMP15	HYPOX	TACHYP
ABC-001	xxxx-xx-xx	xxxx-xx-xx	...	xxxx-xx-xx	>=1800m	Y	Y	...			40
ABC-001	xxxx-xx-xx	xxxx-xx-xx	...	xxxx-xx-xx	>=1800m		Y	...		91	
ABC-001	xxxx-xx-xx	xxxx-xx-xx	...	xxxx-xx-xx	>=1800m	Y		...	Y		50
ABC-001	xxxx-xx-xx	xxxx-xx-xx	...	xxxx-xx-xx	>=1800m			...		90	
ABC-002	xxxx-xx-xx	xxxx-xx-xx	...	xxxx-xx-xx	<1800m	Y		...			45
ABC-002	xxxx-xx-xx	xxxx-xx-xx	...	xxxx-xx-xx	<1800m		Y	...		92	
ABC-002	xxxx-xx-xx	xxxx-xx-xx	...	xxxx-xx-xx	<1800m	Y		...	Y		50
ABC-002	xxxx-xx-xx	xxxx-xx-xx	...	xxxx-xx-xx	<1800m			...		93	

Figure 3: Intermediate dataset snapshot with flags for efficacy endpoints

USUBJID	CESTDT	CEENDT	...	EDPT01	EDPT02	EDPT03	EDPT04	...	EDPNT08	EDPNT09	EDPNT10
ABC-001	xxxx-xx-xx	xxxx-xx-xx	...	Y	N	Y	Y	...	Y	Y	Y
ABC-001	xxxx-xx-xx	xxxx-xx-xx	...	Y	N	N	Y	...	Y	Y	Y
ABC-001	xxxx-xx-xx	xxxx-xx-xx	...	N	N	Y	N	...	Y	N	N
ABC-001	xxxx-xx-xx	xxxx-xx-xx	...	N	Y	N	Y	...	Y	N	Y
ABC-002	xxxx-xx-xx	xxxx-xx-xx	...	Y	N	Y	N	...	N	N	N
ABC-002	xxxx-xx-xx	xxxx-xx-xx	...	N	N	N	Y	...	Y	Y	Y
ABC-002	xxxx-xx-xx	xxxx-xx-xx	...	Y	Y	Y	N	...	N	N	N
ABC-002	xxxx-xx-xx	xxxx-xx-xx	...	Y	N	N	N	...	N	N	N

EFFICACY DATASET (ADEFF) DESIGN

The ADaM Basic Data Structure (BDS) is used for the analysis dataset that contains the efficacy endpoints. The Efficacy Analysis Dataset stores derived endpoints with case/non-case status. The

Figure 4 illustrates ADEFF Metadata information. The clinical endpoints were derived using a complex algorithm for determining efficacy cases, which involves comparing dates and selecting the necessary symptoms, signs, and RT-PCR testing results from the Microbiology Specimen (MB domain). The Figure 5 illustrates ADEFF data snapshot. It was derived with RT-PCR testing results to determine if swabs were obtained within the appropriate window, and to determine for each assessment if it met case definition for each efficacy endpoint. Approximately 30 efficacy endpoints were programmed and included for FDA submission. Additional efficacy endpoints were quickly added per FDA adhoc requests, relying on the flexible intermediate ADEFF dataset.

Figure 4: ADEFF Dataset Metadata

Dataset	Description	Class	Structure	Key Variables	Location	Selection Criteria
ADEFF	Efficacy Analysis Dataset	BASIC DATA STRUCTURE	One record per subject, per parameter (endpoint), per RSV type	USUBJID, PARAMN, RSVTYPE	ADEFF.xpt	All randomized subjects with available measurements.

Figure 5: ADEFF dataset snapshot

STUDYID	USUBJID	...	ADT	PARAM	PARAMCD	RSVTYPE	AVAL	...
ABC	ABC-001	...	xxxx-xx-xx	Efficacy endpoint 1	ENDPNT01	A	1	...
ABC	ABC-001	...	xxxx-xx-xx	Efficacy endpoint 1	ENDPNT01	B	0	...
ABC	ABC-001	...	xxxx-xx-xx	Efficacy endpoint 2	ENDPNT02	A	1	...
ABC	ABC-001	...	xxxx-xx-xx	Efficacy endpoint 2	ENDPNT02	B	0	...
ABC	ABC-001	...	xxxx-xx-xx	Efficacy endpoint 3	ENDPNT03	A	0	...
ABC	ABC-001	...	xxxx-xx-xx	Efficacy endpoint 3	ENDPNT03	B	1	...

ENDPOINT DERIVATION IN ADEFF BY USING INERMEDIATE DATASET

The case determination of efficacy endpoint was based on several dates, availability of events, onset of required symptoms/signs, and RT-PCR testing results.

The

Figure 6 illustrates a simplified example of the efficacy endpoint derivations for ADEFF dataset.

The first endpoint (PARAMCD = ”ENDPNT01”) derived in Figure 5 by selection following elements from intermediate dataset:

EDPT01 = ”Y” (symptoms flag for Efficacy endpoint 1 / ENDPNT01)

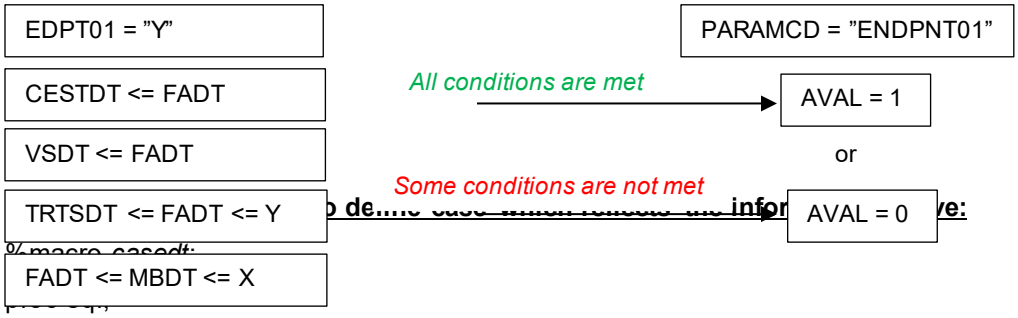
CESTDT <= FADT, VSDT <= FADT (symptom onset or worsening date on or after clinical episode started)

TRTSDT <= FADT <= Y (symptom on or after first dose and less than Y days after first dose)

FADT <= MBDT <= X (RT-PCR positive testing result collected within X days of symptom onset or worsening date)

If all these conditions are met, then assign AVAL = 1 for this endpoint (PARAMCD = ”ENDPNT01”). It will be defined as the case for this efficacy endpoint.

Figure 6: Efficacy endpoint derivation




```
%macro casedt;
        select * into :mb1cnt from (select count(*) as cnt from mb2pos_ group by usubjid) having
cnt=max(cnt);
quit;
data &prmcd._0;
  merge adeff02(in=in1) mb2tr_pos;
        by usubjid;
        if in1;
run;
data &prmcd._1;
  set &prmcd._0;
        %do d=1 %to &mb1cnt;
          if nmiss(cestdt,casedt_,mbdt&d)=0 and max(casedt_-7,cestdt)<=mbdt&d<=casedt_+12 then
                do;
                        if cestdt<=&trtdt or .Z<mbdt&d<=&trtdt then delete;
                        casefl  ='Y';
            if casefl='Y';
            format mbdt casembdt casedt yymmdd10.;
                        output &prmcd._1;
                end;
        %end;
run;
%mend casedt;
```

ABBREVIATIONS

Abbreviation	Expanded Term	Abbreviation	Expanded Term
ADaM	Analysis Data Model	ADSL	Subject-Level Analysis Dataset
SDTM	Study Data Tabulation Model	AE	Adverse Events domain
BDS	Basic Data Structure	CE	Clinical Events domain

RT-PCR	Real Time-Polymerase Chain Reaction	EX	Exposure domain
FDA	Food and Drug Administration	FA	Findings About domain
ADEFF	Efficacy Analysis Dataset	HO	Healthcare Encounters domain
ADVS	Vital Sign Analysis Dataset	MB	Microbiology Specimen domain

CONCLUSION

This paper demonstrates an example where the utilization of an intermediate dataset facilitates the straightforward customization of endpoints for efficacy dataset. A key advantage of this approach lies in its ability to consolidate all essential symptoms, signs, severity indicators, and dates into a single intermediate dataset, thereby enabling the most efficient, rapid, and traceable derivation of endpoints. This capability allows programmers to generate efficacy endpoints in a standardized and consistent manner, ensuring both accuracy and operational efficiency.

REFERENCES

Analysis Data Model ADaM v2.1 (www.cdisc.org)

Analysis Data Model Implementation Guide ADaM IG v1.1 (www.cdisc.org)

Analysis Data Model Basic Data Structure for Time-to-Event (TTE) Analyses (ADaM BDS v1.0) (www.cdisc.org)

Analysis Data Model Structure for Occurrence Data (ADaM OCCDS v1.0) (www.cdisc.org)

Study Data Tabulation Model (SDTM v1.4) (www.cdisc.org)
Study Data Tabulation Model Implementation Guide (SDTM IG v3.2) (www.cdisc.org)

ACKNOWLEDGEMENT

Authors would like to acknowledge Late-Stage Statistical Programming Leadership Team for providing support, opportunity, reviewing and providing important feedback for this paper.

CONTACT INFORMATION

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

Edward Boykis
Merck & Co., Inc., Rahway, NJ, USA
edward.boykis@merck.com

Ying Zhang
Merck & Co., Inc., Rahway, NJ, USA
ying.zhang30@merck.com

Ziqiang Chen
Merck & Co., Inc., Rahway, NJ, USA
ziqiang.chen@merck.com

Peng Wan
Merck & Co., Inc., Rahway, NJ, USA
peng.wan@merck.com