Paper DS01

Deriving Phantom Records in BDS Datasets to Support FDA's Standard Safety Figures for Missing Data Analysis

Ruchitbhai Patel, Merck & Co., Inc., Rahway, NJ, USA
Guowei Wu, Merck & Co., Inc., Rahway, NJ, USA

ABSTRACT

Missing and existing data analyses are key components of safety evaluations for laboratory parameters in NDAs and BLAs, as outlined in recent FDA guidance on standard safety tables and figures. Handling missing data is a critical challenge in clinical trials, potentially impacting study validity. The Basic Data Structure (BDS) dataset, a core part of CDISC ADaM standards, provides a robust framework for creating analysis-ready datasets. This paper presents a systematic approach to derive phantom records within BDS datasets to represent missing observations implied by study design or visit schedules. Incorporating these phantom records enables analysts to better characterize missing data patterns and conduct sensitivity analyses. The method leverages protocol-defined visit windows and data relationships to infer and generate phantom records. Practical examples illustrate how this approach supports missing data assessment and the generation of FDA's standard safety figures, advancing clinical data standards and enhancing the reliability of clinical trial analyses.

INTRODUCTION

Recent FDA guidance on standard safety tables and figures (SSTF) underscores the importance of assessing impact of missing data when evaluating laboratory parameters in NDAs and BLAs. Missing or uncollected assessments can distort safety summaries by introducing bias, reducing statistical power, and obscuring clinically relevant signals. The risks are greater when the set of expected observations is not explicitly defined. The Basic Data Structure (BDS) within the CDISC ADaM framework is the standard vehicle for assembling analysis ready laboratory data for safety outputs. However, analysis datasets based on BDS typically contains only recorded assessments and therefore does not indicate where protocol defined scheduled assessments were expected but not captured. This lack of information related to missing data in BDS hinders generation of FDA aligned safety figures for missing data analyses and reduces transparency for reviewers.

To address these challenges, this paper describes a practical method to derive PHANTOM records in BDS that explicitly marks missing assessments at the subject, parameter, and visit level. This will be illustrated using ADLB dataset with only needed variables. Leveraging subject-level anchors such as randomization and study reference end date, a parameter specific visit schedule, and pre-defined decision rules, the approach creates a subject by visit skeleton of expected observations and populates missing data rows with PHANTOM records. These records maintain a clear separation between observed, absent and imputed data, thereby enabling robust missing data diagnostics and sensitivity analyses. This method improves traceability, facilitates consistent generation of FDA's standard safety figures, and enhances the transparency and integrity of regulatory safety reporting. The remainder of the paper details the derivation logic, implementation considerations with examples.

SUBJECT LEVEL DATA

We leverage the ADSL dataset and other relevant ADaM resources to identify two subject-level anchor dates: the study start date and the study reference end date. For illustration, we use randomization date (`RANDDT`) as the study start and end-of-study date (`EOSDT`) as the study reference end date. Taking the difference `EOSDT`- `RANDDT` + 1, we derive the subject-specific relative study end day (`EOSDY`), which indicates the last day the subject is considered on study for analysis purposes. Depending on study design and available source data, alternative rules may be used to define the subjects relative study end day - for example, date of last on-study assessment, date of discontinuation for any cause, last dose date plus a predefined exposure window, or end of follow-up. Table 1 shows example subject-level anchor

dates and the derived EOSDY. All subjects in Table 1 are assumed to have completed the study and therefore have an observed EOSDT. The example dataset in Table 1 will be used for further illustration in subsequent sections.

USUBJID	RANDDT	EOSDT	EOSDY
1	2024-06-19	2024-07-25	37
2	2024-07-03	2024-08-28	57
3	2024-08-23	2024-09-05	14
4	2024-06-15	2024-07-04	20
5	2024-07-06	2024-08-09	35
6	2024-08-12	2024-09-05	25

Table 1: Example subject-level dataset showing subjects relative study end day (EOSDY) calculated as EOSDT – RANDDT + 1, where RANDDT is the randomization date and EOSDT is study reference end date.

ANALYSIS VISIT SCHEDULE

For each parameter, and guided by the study design, we define visit windows that specify when an assessment is expected to occur. The analysis visit schedule dataset captures these windows and provides, at minimum, the planned visit target day and the window bounds (visit lower bound day and visit upper bound day) relative to the study start. These fields enable standardized mapping of actual assessment dates to planned visits: an assessment falling between the visit upper bound day and visit lower bound day is considered to correspond to that scheduled visit, while assessments outside these bounds may be classified as unscheduled or mapped according to pre-specified rules. The schedule should be documented for each analysis parameter (or parameter group) and include any special-case definitions e.g., multiple expected assessments within a single visit, visit windows that vary by cohort, or windows tied to dosing milestones. Clear definitions of the analysis visit schedule ensure reproducible derivation of visit assignments and support consistent generation of PHANTOM records when expected assessments are not observed. Below Table 2 (Analysis Visit Schedule) lists the planned visits and their windows for PARAM1 used in our examples. Each row shows the visit name (AVISIT), numeric visit order (AVISITN), the analysis parameter (PARAM), and the window bounds expressed as days relative to study start: LOWER_BOUND, TARGET_DAY, and UPPER_BOUND.

AVISIT	AVISITN	PARAM	LOWER_BOUND	TARGET_DAY	UPPER_BOUND
Baseline	1	PARAM1	-1	0	1
Visit 2	2	PARAM1	2	22	31
Visit 3	3	PARAM1	32	43	52
Visit 4	4	PARAM1	53	64	73

Table 2: Planned Visit Windows for parameter "PARAM1" (LOWER_BOUND, TARGET_DAY, UPPER_BOUND day relative to Study Start)

The schedule defines the expected visit windows used to map actual assessment dates to planned visits and to identify expected-but-missing assessments for PHANTOM record derivation. This visit schedule is used to determine which visits are expected per subject (given their EOSDY) and to generate PHANTOM records for any expected but unobserved assessments.

EXPECTED ANALYSIS VISITS

We merge subject-level anchor records from Table 1 (i.e. `RANDDT`, `EOSDT`, and derived `EOSDY`) with the parameter-specific analysis visit schedule from Table 2 to construct a subject by visit skeleton of expected assessments. For each subject, we compare the subject's relative study end day (`EOSDY`) with the scheduled visit window bounds. In the primary simplified rule implemented here, all the scheduled visit whose window lower bound day is less than or equal to subject's relative study end day (i.e. `LOWER_BOUND <= EOSDY`) are designed as expected visits for that subject. An expected visit for which no observed assessment exists is represented by a PHANTOM record in the BDS/ADLB dataset to explicitly mark the missing observation. This rule offers a clear, reproducible approach to identifying expected analysis visits while remaining extensible: study-specific enhancements may incorporate additional decision criteria (for example, discontinuation reason, per-visit eligibility flags, or censoring rules tied to exposure). All applied decision rules and any deviations from the primary rule should be explicitly documented to preserve traceability and to facilitate review and reproducibility of downstream safety summaries and missing-data analyses.

USUBJID	AVISIT	AVISITN	PARAM	LOWER BOUND	UPPER BOUND	RANDDT	EOSDT	EOSDY	Must Have Record
1	Baseline	1	PARAM1	-1	1	2024-06-19	2024-07-25	37	**Y**
1	Visit 2	2	PARAM1	2	31	2024-06-19	2024-07-25	37	**Y**
1	Visit 3	3	PARAM1	32	52	2024-06-19	2024-07-25	37	**Y**
1	Visit 4	4	PARAM1	53	73	2024-06-19	2024-07-25	37	N
2	Baseline	1	PARAM1	-1	1	2024-07-03	2024-08-28	57	**Y**
2	Visit 2	2	PARAM1	2	31	2024-07-03	2024-08-28	57	**Y**
2	Visit 3	3	PARAM1	32	52	2024-07-03	2024-08-28	57	**Y**
2	Visit 4	4	PARAM1	53	73	2024-07-03	2024-08-28	57	**Y**
3	Baseline	1	PARAM1	-1	1	2024-08-23	2024-09-05	14	**Y**
3	Visit 2	2	PARAM1	2	31	2024-08-23	2024-09-05	14	**Y**
3	Visit 3	3	PARAM1	32	52	2024-08-23	2024-09-05	14	N
3	Visit 4	4	PARAM1	53	73	2024-08-23	2024-09-05	14	N
4	Baseline	1	PARAM1	-1	1	2024-06-15	2024-07-04	20	**Y**
4	Visit 2	2	PARAM1	2	31	2024-06-15	2024-07-04	20	**Y**
4	Visit 3	3	PARAM1	32	52	2024-06-15	2024-07-04	20	N
4	Visit 4	4	PARAM1	53	73	2024-06-15	2024-07-04	20	N
5	Baseline	1	PARAM1	-1	1	2024-07-06	2024-08-09	35	**Y**
5	Visit 2	2	PARAM1	2	31	2024-07-06	2024-08-09	35	**Y**
5	Visit 3	3	PARAM1	32	52	2024-07-06	2024-08-09	35	**Y**
5	Visit 4	4	PARAM1	53	73	2024-07-06	2024-08-09	35	N
6	Baseline	1	PARAM1	-1	1	2024-08-12	2024-09-05	25	**Y**
6	Visit 2	2	PARAM1	2	31	2024-08-12	2024-09-05	25	**Y**
6	Visit 3	3	PARAM1	32	52	2024-08-12	2024-09-05	25	N
6	Visit 4	4	PARAM1	53	73	2024-08-12	2024-09-05	25	N

Table 3: Example dataset obtained by combining subject-level anchor records from Table 1 and the parameter specific analysis visit schedule from Table 2.

Table 3 illustrates the merged subject-by-visit skeleton used to identify expected analysis visits for PARAM1. The `Must Have Record` flag indicates whether the visit is expected for that subject under the simplified rule used here (a visit is expected if its window lower bound day is on or before subject relative study end day i.e. `LOWER_BOUND <= EOSDY`). Rows with `Must Have Record = 'Y'` correspond to visits for which an observed assessment should exist (otherwise a PHANTOM record will be created); rows with `Must Have Record = 'N'` indicate scheduled visits that fall after

the subject's EOSDY and are therefore not expected for that subject. This skeleton is used to identify where PHANTOM records should be created when an expected observation is not present in the analysis dataset. From the Table 3 above for PARAM = "PARAM1" below set of subjects and analysis visit should be present in BDS (ADLB for our example purpose) dataset.

- USUBJID 1 should have records for AVISIT 1 to 3.
- USUBJID 2 should have records for AVISIT 1 to 4.
- USUBJID 3 should have records for AVISIT 1 and 2.
- USUBJID 4 should have records for AVISIT 1 and 2.
- USUBJID 5 should have records for AVISIT 1 to 3.
- USUBJID 6 should have records for AVISIT 1 and 2.

ADD "PHANTOM" RECORDS

Based on the subject-by-visit skeleton in Table 3, we compare the expected analysis visits for a given parameter (e.g., PARAM = "PARAM1") with the actual ADLB records to identify scheduled visits that lack laboratory observations. For each such expected-but-missing visit, we add a new ADLB row with DTYPE = "PHANTOM" and set AVAL to missing. The PHANTOM record is populated with the planned visit identifiers (AVISIT, AVISITN) and the parameter identifiers (e.g., PARAM, PARAMCD), and includes the subject anchors (RANDDT, EOSDT) and derived EOSDY for context. Using DTYPE = "PHANTOM" follows CDISC terminology and preserves a clear distinction between explicitly missing expected assessments and any imputed values, ensuring downstream analyses and regulatory outputs handle these rows appropriately. This concise, per-subject augmentation provides an auditable trail of visit-level gaps flagged by the derivation logic and demonstrates how the ADLB is augmented to explicitly represent missing observations. Detailed examples of the inserted PHANTOM rows and their associated visit and subject identifiers are shown in Table 4 for PARAM = "PARAM1". Summary of PHANTOM records shown in Table 4 is provided below. Note that no PHANTOM records were added for USUBJID 3 because ADLB already contained all expected visit records for that subject.

- USUBJID 1 PHANTOM record for AVISIT = 3 added to ADLB.
- USUBJID 2 PHANTOM record for AVISIT = 2 added to ADLB.
- USUBJID 3 No PHNATOM records added.
- USUBJID 4 PHANTOM record for AVISIT = 2 added to ADLB.
- USUBJID 5 PHANTOM record for AVISIT = 2 added to ADLB.
- USUBJID 6 PHANTOM record for AVISIT = 2 added to ADLB.

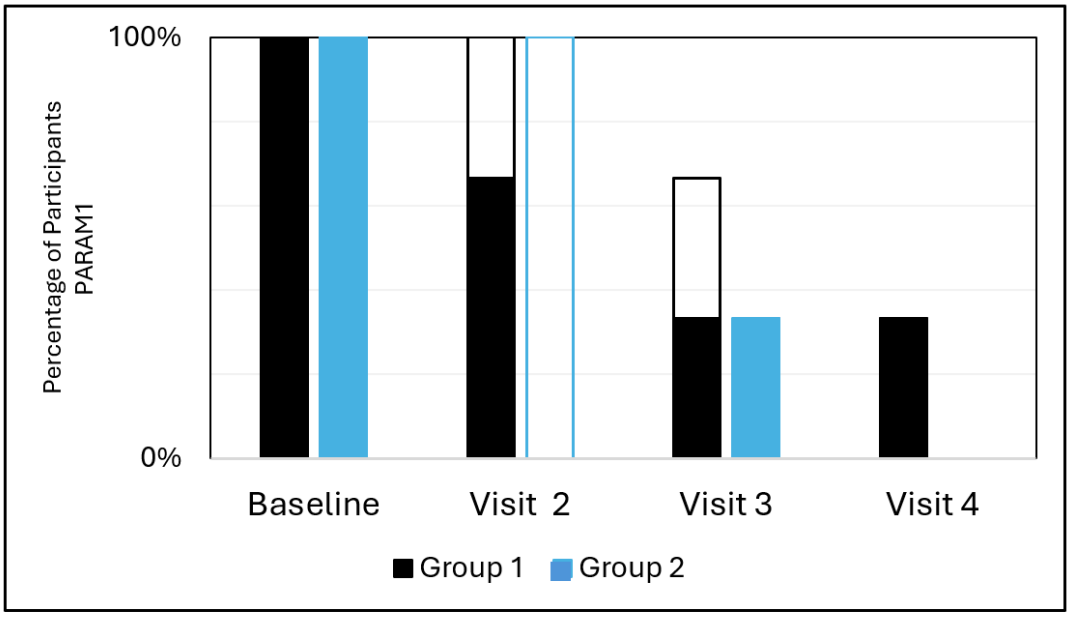
USUBJID	DOMAIN	AVISIT	AVISITN	TRT0XP	PARAMCD	AVAL	DTYPE
1	ADLB	Baseline	1	**Group 1**	PARAM1	x.x	
1	ADLB	Visit 2	2	**Group 1**	PARAM1	x.x	
1	ADLB	Visit 3	3	**Group 1**	PARAM1		PHANTOM
2	ADLB	Baseline	1	**Group 1**	PARAM1	x.x	
2	ADLB	Visit 2	2	**Group 1**	PARAM1		PHANTOM
2	ADLB	Visit 3	3	**Group 1**	PARAM1	x.x	
2	ADLB	Visit 4	4	**Group 1**	PARAM1	x.x	
3	ADLB	Baseline	1	**Group 1**	PARAM1	x.x	
3	ADLB	Visit 2	2	**Group 1**	PARAM1	x.x	
4	ADLB	Baseline	1	**Group 2**	PARAM1	x.x	
4	ADLB	Visit 2	2	**Group 2**	PARAM1		PHANTOM
5	ADLB	Baseline	1	**Group 2**	PARAM1	x.x	
5	ADLB	Visit 2	2	**Group 2**	PARAM1		PHANTOM
5	ADLB	Visit 3	3	**Group 2**	PARAM1	x.x	
6	ADLB	Baseline	1	**Group 2**	PARAM1	x.x	
6	ADLB	Visit 2	2	**Group 2**	PARAM1		PHANTOM

Table 4: ADLB Records with Inserted PHANTOM Rows for PARAM1. Showing observed and PHANTOM entries by subject, visit, treatment group, and parameter code.

ANALYSIS

Using the augmented ADLB (observed records plus derived PHANTOM rows), we make expected observations explicit at the subject, by-treatment, by-parameter, and by-visit level, enabling comprehensive and traceable missing-data analyses. Figure 1 is produced from this augmented ADLB. In this example, both Group 1 and Group 2 include three participants each. For each treatment group and planned visit, Figure 1 shows:

- The observed percentage (solid portion of the bar) of participants with an actual recorded result at that visit in the analysis population for the treatment group.
    - Observed percentage = (no. of participants with results / analysis population) x 100
- The expected percentage (bar height) of participants remaining in the trial at that visit, defined as observed + PHANTOM records at that visit in the analysis population for the treatment group.
    - Expected percentage = (no. of participants remaining in trial / analysis population) × 100.
- The inset table - formatted as “observed counts / expected counts” (no. of participants with results / no. of participants remaining in trial), providing the numeric counts that show where expected assessments were uncollected.



No. of Participants: Remaining in Trial with Results (Solid Bar) / Remaining in Trial (Bar Height)

Treatment	Baseline	Visit 2	Visit 3	Visit 4
Group 1	3/3	2/3	1/2	1/1
Group 2	3/3	0/3	1/1	0/0

Figure 1: Proportion of participants with observed and expected counts for PARAM1 results by visit and treatment group.

By explicitly including PHANTOM records in the denominator, the figure clarifies which participants were expected to be assessed at each visit and provides transparent, reproducible percentages. Because PHANTOM rows represent only expected-but-uncollected assessments and are not imputations, reviewers can clearly distinguish missingness from observed data. Overall, Figure 1 demonstrates that augmenting ADLB with PHANTOM records improves

denominator definition and supports transparent missing-data assessment and reporting consistent with FDA standard safety analyses.

CONCLUSION

Deriving PHANTOM records within BDS/ADLB datasets provides a practical, auditable way to make expected-but-uncollected assessments explicit, improving transparency and reproducibility of safety analyses. By leveraging subject-level anchors (e.g., randomization and study-end dates), analysis visit schedule, and clear derivation rules, PHANTOM records enable unambiguous definitions, more informative missing-data diagnostics, and structured sensitivity analyses without conflating absence with imputation. This approach supports consistent generation of FDA-aligned safety tables and figures, facilitates reviewer inspection, and strengthens the integrity of regulatory submissions. Future work should focus on standardizing derivation rules across studies and automating implementation to ensure broad, consistent adoption.

REFERENCES

CDISC ADaM Controlled Terminology, 2025-09-26, from https://evs.nci.nih.gov/ : https://evs.nci.nih.gov/ftp1/CDISC/ADaM/ADaM%20Terminology.pdf

FDA U.S. Food & Drug Administration, Standard Safety Table and Figures: Integrated Guide 2025-04-01, from https://www.fda.gov/ : https://www.fda.gov/media/187065/download

ACKNOWLEDGEMENT

The authors would like to thank management teams from Merck & Co., Inc. Rahway, NJ, USA, for their advice and this paper/presentation.

CONTACT INFORMATION

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

Author Name: Ruchitbhai Patel
Company: Merck & Co., Inc., Rahway, NJ, USA
Phone: 267-305-1417
Email: ruchitbhai.patel@merck.com

Author Name: Guowei Wu
Company: Merck & Co., Inc., Rahway, NJ, USA
Phone: 267-305-5381
Email: guowei_wu@merck.com