PharmaSUG 2023 - Paper DS-051

The Phantom of the ADaM: Adding Missing Records to BDS Datasets

Anastasiia Drach, Intego Group

ABSTRACT

Missing data is a 'pain' of any study. There are many imputation techniques available, but sometimes all we need to know is just that the data is missing. In these cases, it is useful to add derived records to your ADaM datasets with missing AVAL/AVALC to indicate missed visits or timepoints. Such records are called phantom records. In this paper, we discuss how to add them into BDS ADaM dataset using PRO data as an example. We will start with an overview of different ways to represent missing data in SDTM. The paper will present several types of analysis which require the inclusion of phantom records to account for missing data. It will cover various scenarios of adding such records, from the most straightforward to more complex ones. Finally, we will provide some ready-to-use solutions for the creation of phantom records, which could be easily adjusted to your individual needs.

INTRODUCTION

Missing data is a common challenge that arises in clinical trials, which can impact the validity and reliability of the study results. The reasons for missing data can vary, including missed visits, participant dropout, incomplete data collection, or technical issues with data recording. Regardless of the cause, missing data can create issues in the analysis and that is why it is important to learn to deal with it.

To address the issue of missing data in clinical trials, CDISC has developed standards for capturing and handling missing data. These standards are designed to promote transparency and consistency in the analysis of clinical trial data. SDTM standard includes recommendations on how to present the missing data depending on the way it was collected (if at all). ADaM guidelines facilitate the handling of missing data that is more suitable for analysis purposes, including the addition of derived records using various imputation techniques. While imputation is an indispensable tool in many cases, quite often it is sufficient to know when and which particular data is missing and to have a record with a missing value in the dataset for each parameter to represent it. Unfortunately, this is not always the case with the data collection, some missing data is often just absent from the dataset without any trace of it, or present in the way that does not work for your analysis needs.

Luckily, CDISC offers a solution for such cases as well. BDS analysis datasets are allowed to have derived observations, including the ones which are imputed, with a DTYPE variable containing the derivation method used to create such record. This variable has a list of available controlled terminology values, one of them being 'PHANTOM'. Phantom record imputation technique is a technique that creates a record with a missing analysis value when there is no observed record for a given analysis visit or analysis timepoint (see [1] , p.9).

This paper will cover in detail various scenarios of when phantom records can be created, different ways of adding such records to BDS ADaM datasets depending on the way your data is collected and represented in SDTM and the types of analyses where the presence of phantom records would be required.

COMMON USE CASES FOR PHANTOM RECORDS

Phantom records are designed to help you represent and analyze the data that is missing when it was expected to be collected. For example, let's say a patient is expected to come to the clinic every 21 days for a certain assessment and the corresponding visits are Cycle 1, Cycle 2, etc. The patient comes at Cycle 1, misses Cycle 2, then goes on with the rest of the visits as scheduled until discontinuing study after Cycle 7. In this case, at Cycle 2 the patient is expected to have a record, but doesn't, whereas Cycles from 8 onwards are no longer expected, since the subject discontinued after Cycle 7. An End of treatment visit would be expected as well. So, we would create a phantom record for Cycle 2 and End of treatment visit to be able to show and analyze the fact that these assessments were missed, however we would not be adding any cycles past Cycle 7.

There are many scenarios where the analysis that needs to be performed may require the creation of phantom records in your ADaM datasets. Let's take a look at the most common examples:

- **Exposure analysis:** when the summary of number of missed doses is required, one of the easiest ways to approach it is to add phantom records to your exposure dataset. In exposure data more often than not missed doses would be still collected on CRF along with the reason why they were missed. However, in some cases if the subject just skipped the whole visit altogether, it is possible that there would be no record for it. In that case, unless you create a phantom record for such a visit, the number of missed doses may be calculated incorrectly very easily.
- **Efficacy analysis:** here missing data is especially critical. It is quite common to perform certain sensitivity analyses to assess the potential impact of missing data and to adjust censoring rules in cases where an event occurred after 1 or 2 or more missed assessments. Hence, it is important to know exactly which assessments a patient was expected to complete but did not, in order to apply censoring rules correctly.
- **PRO analysis:** the analysis that is, probably, suffering the most from the problem of missing data is PRO analysis and that is why we will use it as an example case throughout the paper. So let's discuss in more detail what is PRO data and when we would need to add phantom records to our PRO datasets.

OVERVIEW OF PRO DATA

PRO stands for Patient-Reported Outcome, which is a type of data collected directly from patients about their health and how they are feeling. It is a way to measure a patient's subjective experience of their disease, symptoms, and treatment. PRO data is usually collected through questionnaires that patients fill out themselves. These questionnaires may ask about patient's symptoms, their ability to perform certain tasks or activities, their quality of life, and their satisfaction with their treatment. The data collected from these questionnaires can be used to assess the efficacy and safety of a particular treatment, to understand how patient's disease affects their daily life, and to evaluate the impact of treatment on a patient's quality of life.

A common issue that can arise with PRO data in clinical trials is missing data. Missing data occurs when a patient does not complete all of the questions on a questionnaire, or when a patient drops out of a trial before completing all of the PRO assessments. Since patients complete these questionnaires on their own, missing questions or even whole assessments occur more often than with any other type of data collection. This may be even more prominent with digital solutions that are becoming more popular in the recent years. Hence it is important to keep track of it and take it into account.

EORTC-QLQ QUESTIONNAIRE

There is a large variety of different questionnaires in clinical trials, but one of the most widely used ones in oncology is the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire. The EORTC-QLQ questionnaire is a tool used to measure the impact of the disease and its treatment on the quality of life of cancer patients participating in clinical trials. The core questionnaire includes 30 questions that cover a wide range of topics, such as physical functioning, pain, fatigue, emotional functioning, and social functioning. The EORTC-QLQ questionnaire groups questions into multiple scales, and then the score for each scale is calculated using a specifically designed formula based on the individual question scores and on the number of questions that contribute to each score. The questionnaire consists of functional scales, symptom scales/single items and global health status (see [2]):

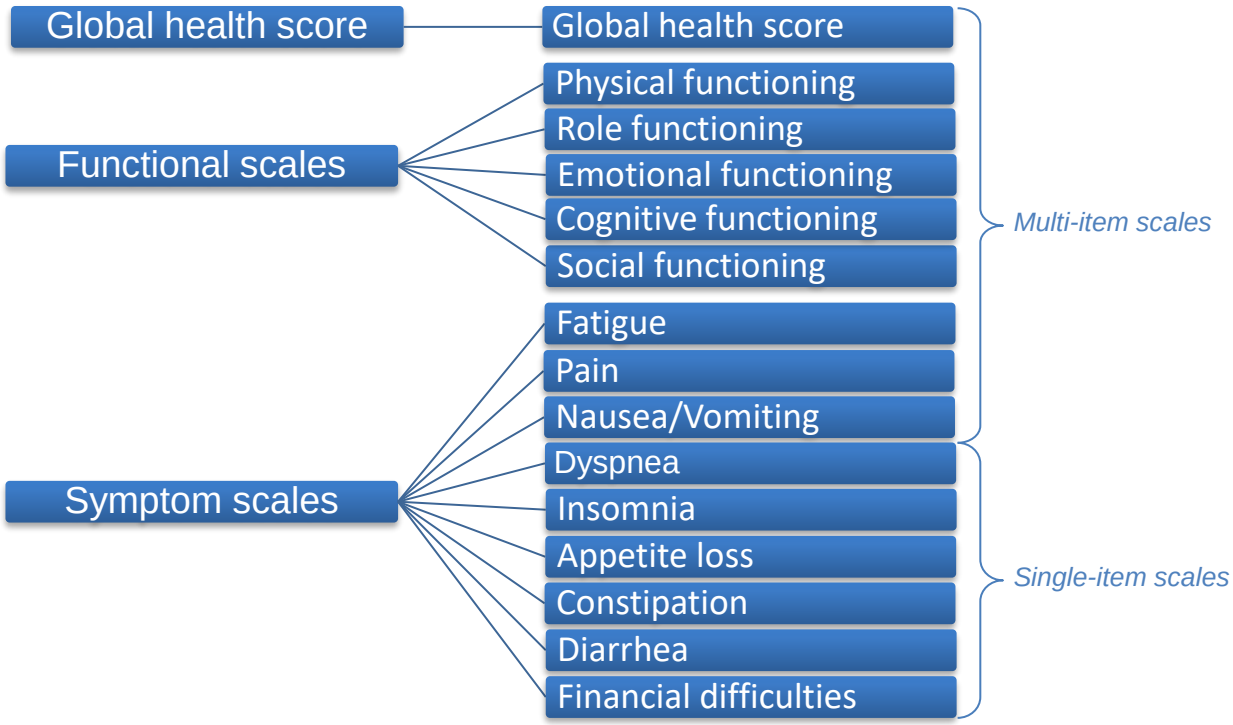

Figure 1. EORTC-QLQ Scales

REPRESENTATION IN AN ADAM DATASET

Below is an example of a standard SDTM.QS dataset containing EORTC-QLQ data with no missed visits/assessments for now (for demonstration purposes only selected variables are kept):

	DOMAIN	USUBJID	QSTESTCD	QSSTRESN	QSBLFL	VISITNUM	VISIT
1	QS	XXX-001-10101	QLQ01	1	Y	2	CYCLE 1 DAY 1
2	QS	XXX-001-10101	QLQ01	2		3	CYCLE 2 DAY 1
3	QS	XXX-001-10101	QLQ01	1		4	CYCLE 3 DAY 1
4	QS	XXX-001-10101	QLQ01	1		5	CYCLE 4 DAY 1
5	QS	XXX-001-10101	QLQ01	4		8000	END OF TREATMENT
6	QS	XXX-001-10101	QLQ02	3	Y	2	CYCLE 1 DAY 1
7	QS	XXX-001-10101	QLQ02	2		3	CYCLE 2 DAY 1
8	QS	XXX-001-10101	QLQ02	3		4	CYCLE 3 DAY 1
9	QS	XXX-001-10101	QLQ02	3		5	CYCLE 4 DAY 1
10	QS	XXX-001-10101	QLQ02	4		8000	END OF TREATMENT
11	QS	XXX-001-10101	QLQ03	3	Y	2	CYCLE 1 DAY 1
12	QS	XXX-001-10101	QLQ03	2		3	CYCLE 2 DAY 1
13	QS	XXX-001-10101	QLQ03	1		4	CYCLE 3 DAY 1
14	QS	XXX-001-10101	QLQ03	1		5	CYCLE 4 DAY 1
15	QS	XXX-001-10101	QLQ03	2		8000	END OF TREATMENT

Display 1. Example of SDTM.QS dataset with no missing data

In a corresponding ADaM dataset (let's call it ADQS) we will keep the source records from SDTM for traceability, as well as add new derived parameters (PARAMTYP = 'DERIVED') to calculate a summary score for each of the scales in EORTC-QLQ questionnaire. Since we don't have any missing data in this example, ADQS can be generated with no additional steps:

	USUBJID	AVISIT	AVISITN	PARAM	PARAMCD	PARAMTYP	DTYPE	AVAL	ABLFL
148	XXX-001-10101	BASELINE	1	Physical Functioning Score	QLQPFSC	DERIVED		73	Y
149	XXX-001-10101	CYCLE 2 DAY 1	3	Physical Functioning Score	QLQPFSC	DERIVED		80	
150	XXX-001-10101	CYCLE 3 DAY 1	4	Physical Functioning Score	QLQPFSC	DERIVED		80	
151	XXX-001-10101	CYCLE 4 DAY 1	5	Physical Functioning Score	QLQPFSC	DERIVED		87	
152	XXX-001-10101	END OF TREATMENT	8000	Physical Functioning Score	QLQPFSC	DERIVED		53	
158	XXX-001-10101	BASELINE	1	Emotional Functioning Score	QLQEFSC	DERIVED		42	Y
159	XXX-001-10101	CYCLE 2 DAY 1	3	Emotional Functioning Score	QLQEFSC	DERIVED		33	
160	XXX-001-10101	CYCLE 3 DAY 1	4	Emotional Functioning Score	QLQEFSC	DERIVED		58	
161	XXX-001-10101	CYCLE 4 DAY 1	5	Emotional Functioning Score	QLQEFSC	DERIVED		50	
162	XXX-001-10101	END OF TREATMENT	8000	Emotional Functioning Score	QLQEFSC	DERIVED		50	
163	XXX-001-10101	BASELINE	1	Cognitive Functioning Score	QLQCFSC	DERIVED		50	Y
164	XXX-001-10101	CYCLE 2 DAY 1	3	Cognitive Functioning Score	QLQCFSC	DERIVED		50	
165	XXX-001-10101	CYCLE 3 DAY 1	4	Cognitive Functioning Score	QLQCFSC	DERIVED		50	
166	XXX-001-10101	CYCLE 4 DAY 1	5	Cognitive Functioning Score	QLQCFSC	DERIVED		67	
167	XXX-001-10101	END OF TREATMENT	8000	Cognitive Functioning Score	QLQCFSC	DERIVED		33	

Display 2. Example of ADAM.ADQS dataset with no missing data

With all the data available it is pretty straightforward. But how can the missing data affect the analysis that is required and, hence, the structure and approaches to creation of our analysis dataset?

COMPLETION RATE ANALYSIS

The most common types of analysis for PRO data, and EORTC-QLQ questionnaire results in particular, are summary of AVAL and CHG of each scale by visit. There could also be shift tables from baseline to worst post-baseline, improvement or deterioration analysis. Another common example, and the one that analyses missing data, is a completion rate output and below is a sample table that you might need to generate:

Table xx.x.x Summary of Completion Rate of EORTC-QLQ-C30 by Visit

Scale/Item: <……>

Visit		Drug A (N=xx)	Drug B (N=xx)
Baseline	Number of patients expected to complete	xx	xx
	All questions completed	xx (xx.x)	xx (xx.x)
	At least half of the questions completed, but not all	xx (xx.x)	xx (xx.x)
	At least one question completed, but less than half	xx (xx.x)	xx (xx.x)
	None of the questions completed	xx (xx.x)	xx (xx.x)
Cycle 2 Day 1	Number of patients expected to complete	xx	xx
	All questions completed	xx (xx.x)	xx (xx.x)
	At least half of the questions completed, but not all	xx (xx.x)	xx (xx.x)
	At least one question completed, but less than half	xx (xx.x)	xx (xx.x)
	None of the questions completed	xx (xx.x)	xx (xx.x)
Cycle 3 Day 1	Number of patients expected to complete	xx	xx
	All questions completed	xx (xx.x)	xx (xx.x)
	At least half of the questions completed, but not all	xx (xx.x)	xx (xx.x)
	At least one question completed, but less than half	xx (xx.x)	xx (xx.x)
	None of the questions completed	xx (xx.x)	xx (xx.x)

Display 3. Sample completion rate table

For this table you need to know not the actual scores for each scale, but how many questions in each scale have been completed. In order to be able to calculate a score you need to have at least half of the questions in it answered, so the categories analyzed are All questions completed, At least half of the questions completed, but not all, At least one question completed, but less than half or None of the questions completed. To capture these multiple criteria for each parameter it is convenient to utilize MCRIT/MCRITyML set of variables. The first three categories are pretty straightforward and can be derived at the same time as deriving the actual score. The last category, however, that is highlighted in yellow in Display 3, is more tricky. As you can imagine, if none of the questions were completed at a given visit, by default there wouldn't be a record in the dataset, so it would be impossible to have any

records with MCRITyML = 'None of the questions completed'. That was exactly the problem that we had in our experience and the solution to it turned out to be the creation of phantom records.

ADDING PHANTOM RECORDS TO AN ADAM DATASET

The approach that you need to take to add phantom records to your ADaM dataset will heavily depend on the amount and type of data that is missing and on the way such data is captured (or not captured) at SDTM level.

ONLY SOME QUESTIONS ARE MISSING

The first possibility is that a patient had a required visit, completed the questionnaire, but skipped some questions or blocks of questions. There could be several reasons for that:

- *by design:* if one of the questions is answered in a certain way then the consecutive block(s) become irrelevant
- *by accident:* the patient just missed a question or two unintentionally
- *by choice:* the question was uncomfortable or patient refused to answer it for any other reason

What matters to us is whether the question is missing by design or not. If the questionnaire that you are working with has scenarios in which certain questions become irrelevant, you would need to account for that in your code to make sure that you are not adding records that don't make sense. If it is not by design, then you would need to impute such values using phantom records technique.

One thing to remember is that in your completion rate analysis you will be using only the derived records from ADQS dataset, the ones that contain the summary scores for each scale. When you think about that it becomes obvious that it is not necessary to add phantom records for each individual missed question – this will help you to avoid extra work. What is important is to create a phantom record for each missing derived parameter, meaning you will need to create them only in cases when the whole set of questions used to derive a certain scale is missing.

Let's take a look at an example below: a patient did not answer questions 6 and 7 in the questionnaire at Cycle 2 Day 1 visit:

	DOMAIN	USUBJID	QSTESTCD	QSSTRESN	QSBLFL	VISITNUM	VISIT	
21	QS	XXX-001-10101	QLQ05	1	Y	2	CYCLE 1 DAY 1	
22	QS	XXX-001-10101	QLQ05	1		3	CYCLE 2 DAY 1	
23	QS	XXX-001-10101	QLQ05	1		4	CYCLE 3 DAY 1	
24	QS	XXX-001-10101	QLQ05	1		5	CYCLE 4 DAY 1	
25	QS	XXX-001-10101	QLQ05	1		8000	END OF TREATMENT	
26	QS	XXX-001-10101	QLQ06	1	Y	2	CYCLE 1 DAY 1	No CYCLE 2 DAY 1
27	QS	XXX-001-10101	QLQ06	1		4	CYCLE 3 DAY 1	
28	QS	XXX-001-10101	QLQ06	1		5	CYCLE 4 DAY 1	
29	QS	XXX-001-10101	QLQ06	3		8000	END OF TREATMENT	
30	QS	XXX-001-10101	QLQ07	3	Y	2	CYCLE 1 DAY 1	No CYCLE 2 DAY 1
31	QS	XXX-001-10101	QLQ07	3		4	CYCLE 3 DAY 1	
32	QS	XXX-001-10101	QLQ07	1		5	CYCLE 4 DAY 1	
33	QS	XXX-001-10101	QLQ07	4		8000	END OF TREATMENT	
34	QS	XXX-001-10101	QLQ08	1		3	CYCLE 2 DAY 1	
35	QS	XXX-001-10101	QLQ08	1		4	CYCLE 3 DAY 1	
36	QS	XXX-001-10101	QLQ08	1		5	CYCLE 4 DAY 1	
37	QS	XXX-001-10101	QLQ08	2		8000	END OF TREATMENT	

Display 4. Example of SDTM.QS dataset with some data missing

Those are the only two questions that contribute to a Role functioning scale score. Without any imputation the whole derived parameter for this scale (PARAMCD = 'QLQRFSC') will be missing. However, we can add a piece of code to add the missing scales in such cases. We would need to create a dataset containing all of the derived parameters, then create a template with all of these parameters added for all

patients and visits available in the data. Finally, we would merge our main ADQS dataset with this template:

```
data param;
  do paramcd = 'QLQPFSC', 'QLQRFSC', 'QLQEFSC', 'QLQCFSC', 'QLQSFSC',
               'QLQFASC', 'QLQNVSC', 'QLQPASC', 'QLQQLSC';
    paramtyp = 'DERIVED';
    param = put(paramcd, qsparam.);
    output;
  end;
run;

proc sql noprint;
  create table adqs_all (drop = _:) as
  select a.*, b.usubjid, b.avisitn, b.avisit, b.paramcd, b.param, b.paramtyp,
         (case when missing(_usubjid) then 'PHANTOM'
          else '' end) as dtype
  from adqs (rename = (usubjid = _usubjid avisitn = _avisitn avisit = _avisit
            paramcd = _paramcd param = _param) drop = paramtyp dtype) as a
  right join
  (select distinct usubjid, avisitn, avisit, p.paramcd, p.param, p.paramtyp
   from adqs, param as p) as b
  on a._usubjid = b.usubjid and a._avisitn = b.avisitn and
     a._paramcd = b.paramcd;
quit;

data adqs_all;
  set adqs_all;
  if dtype = 'PHANTOM' then do;
    mcrit2ml = 'None of the questions completed';
    mcrit2   = 'EORTC-QLQ Completion Status';
  end;
run;
```

After performing the above mentioned steps you will have successfully added phantom records to ADQS dataset:

	USUBJID	AVISIT	AVISITN	PARAM	PARAMCD	PARAMTYP	DTYPE	AVAL	ABLFL	MCRIT2ML
1	XXX-001-10101	BASELINE	1	Physical Functioning Score	QLQPFSC	DERIVED		73	Y	All questions completed
2	XXX-001-10101	CYCLE 2 DAY 1	3	Physical Functioning Score	QLQPFSC	DERIVED		80		All questions completed
3	XXX-001-10101	CYCLE 3 DAY 1	4	Physical Functioning Score	QLQPFSC	DERIVED		80		All questions completed
4	XXX-001-10101	CYCLE 4 DAY 1	5	Physical Functioning Score	QLQPFSC	DERIVED		87		All questions completed
5	XXX-001-10101	END OF TREATMENT	8000	Physical Functioning Score	QLQPFSC	DERIVED		53		All questions completed
6	XXX-001-10101	BASELINE	1	Role Functioning Score	QLQRFSC	DERIVED		67	Y	All questions completed
7	XXX-001-10101	CYCLE 2 DAY 1	3	Role Functioning Score	QLQRFSC	DERIVED	PHANTOM	.		None of the questions completed
8	XXX-001-10101	CYCLE 3 DAY 1	4	Role Functioning Score	QLQRFSC	DERIVED		67		All questions completed
9	XXX-001-10101	CYCLE 4 DAY 1	5	Role Functioning Score	QLQRFSC	DERIVED		100		All questions completed
10	XXX-001-10101	END OF TREATMENT	8000	Role Functioning Score	QLQRFSC	DERIVED		17		All questions completed
11	XXX-001-10101	BASELINE	1	Emotional Functioning Score	QLQEFSC	DERIVED		42	Y	All questions completed
12	XXX-001-10101	CYCLE 2 DAY 1	3	Emotional Functioning Score	QLQEFSC	DERIVED		33		All questions completed
13	XXX-001-10101	CYCLE 3 DAY 1	4	Emotional Functioning Score	QLQEFSC	DERIVED		58		All questions completed
14	XXX-001-10101	CYCLE 4 DAY 1	5	Emotional Functioning Score	QLQEFSC	DERIVED		50		All questions completed
15	XXX-001-10101	END OF TREATMENT	8000	Emotional Functioning Score	QLQEFSC	DERIVED		50		All questions completed

Display 5. Example of ADaM.ADQS dataset with a phantom record added

A WHOLE VISIT OR TIMEPOINT IS MISSING

We have discussed the possibility of only certain questions being missing at a visit. But what about the cases, when the whole assessment was skipped altogether? We would need to create phantom records with missing values of AVAL for all of the scales in a questionnaire in order to be able to properly represent this fact in a completion rate table, which is summarized by scale and by visit. There are several options of how a missing visit could be captured in SDTM, so let's discuss them one by one to see what difference would it make for us in terms of adding phantom records to ADQS.

QSALL observations

In case a whole questionnaire was not answered at a certain visit, you could, in theory, have a record with missing QSORRES for every single question. However, you can imagine why this would be tedious both to map and to interpret, therefore SDTM has created a dedicated solution for cases like this. If a whole set of assessments was skipped at a certain visit or timepoint, the suggested solution is to create a record with --TESTCD = '--ALL'. Below you can see an example of a patient having missed assessments at Cycles 1 through 3:

	DOMAIN	USUBJID	QSTESTCD	QSSTRESN	QSBLFL	VISITNUM	VISIT
200	QS	XXX-001-10102	QLQ28	2		5	CYCLE 4 DAY 1
201	QS	XXX-001-10102	QLQ28	4		8000	END OF TREATMENT
202	QS	XXX-001-10102	QLQ29	6		5	CYCLE 4 DAY 1
203	QS	XXX-001-10102	QLQ29	4		8000	END OF TREATMENT
204	QS	XXX-001-10102	QLQ30	5		5	CYCLE 4 DAY 1
205	QS	XXX-001-10102	QLQ30	4		8000	END OF TREATMENT
206	QS	XXX-001-10102	QSALL	.		2	CYCLE 1 DAY 1
207	QS	XXX-001-10102	QSALL	.		3	CYCLE 2 DAY 1
208	QS	XXX-001-10102	QSALL	.		4	CYCLE 3 DAY 1

Display 6. Example of SDTM.QS dataset a QSALL record

So, if you are lucky, you will have a QSTESTCD = 'QSALL' in your SDTM.QS dataset and the only thing that you would need to do is to "multiply" it into all the derived scale records that you need.

```
data adqs_ph;
  set adqs;
  if not missing(visitnum) and paramcd ='QSALL' then
  do paramcd = 'QLQPFSC', 'QLQRFSC', 'QLQEFSC', 'QLQCFSC', 'QLQSFSC',
               'QLQFASC', 'QLQNVSC', 'QLQPASC', 'QLQQLSC';
    dtype    = 'PHANTOM';
    mcrit2ml = 'None of the questions completed';
    mcrit2   = 'EORTC-QLQ Completion Status';
    param    = put(paramcd, qsparam.);
    output;
  end;
run;
```

After implementing the step above, ADQS dataset will contain a newly derived phantom record for each of the scales in the questionnaire for all the visits that had a QSALL record in SDTM:

	USUBJID	AVISIT	AVISITN	PARAM	PARAMCD	PARAMTYP	DTYPE	AVAL	ABLFL	MCRIT2ML
86	XXX-001-10102	CYCLE 1 DAY 1	2	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
87	XXX-001-10102	CYCLE 2 DAY 1	3	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
88	XXX-001-10102	CYCLE 3 DAY 1	4	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
89	XXX-001-10102	CYCLE 4 DAY 1	5	Emotional Functioning Score	QLQEFSC	DERIVED		75		All questions completed
90	XXX-001-10102	END OF TREATMENT	8000	Emotional Functioning Score	QLQEFSC	DERIVED		75		All questions completed
91	XXX-001-10102	CYCLE 1 DAY 1	2	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
92	XXX-001-10102	CYCLE 2 DAY 1	3	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
93	XXX-001-10102	CYCLE 3 DAY 1	4	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
94	XXX-001-10102	CYCLE 4 DAY 1	5	Cognitive Functioning Score	QLQCFSC	DERIVED		83		All questions completed
95	XXX-001-10102	END OF TREATMENT	8000	Cognitive Functioning Score	QLQCFSC	DERIVED		50		All questions completed
96	XXX-001-10102	CYCLE 1 DAY 1	2	Social Functioning Score	QLQSFSC	DERIVED	PHANTOM	.		None of the questions completed
97	XXX-001-10102	CYCLE 2 DAY 1	3	Social Functioning Score	QLQSFSC	DERIVED	PHANTOM	.		None of the questions completed
98	XXX-001-10102	CYCLE 3 DAY 1	4	Social Functioning Score	QLQSFSC	DERIVED	PHANTOM	.		None of the questions completed
99	XXX-001-10102	CYCLE 4 DAY 1	5	Social Functioning Score	QLQSFSC	DERIVED		100		All questions completed
100	XXX-001-10102	END OF TREATMENT	8000	Social Functioning Score	QLQSFSC	DERIVED		67		At least half of the questions completed, but not all

Display 7. Example of ADaM.ADQS dataset with a whole visit imputed from a QSALL record

The idea behind these '--ALL' records is to show that a patient was expected to take a questionnaire (or any other test or assessment) at a certain visit, but didn't. This way if this kind of data was first collected and then mapped this way to SDTM, you wouldn't need to check additionally which visits are expected for a certain patient and this specific type of assessment.

Merge with SV or TV

If you are less lucky though, there will be no QSALL records in SDTM.QS, either for some of the visits or in general. Then you would need to perform some additional manipulations to determine which visits are expected for a patient. The way to do that is to merge your QS dataset with SDTM.SV (or in certain cases SDTM.TV) to get the whole list of visits for a patient. This way you will be able to add all the visits that a patient was expected to have without the risk of adding any unnecessary visits after a patient has already discontinued study. One thing to be mindful of is that questionnaires are not necessarily required to be filled out at all visits that a patient can have in SV or TV. Hence, you need to review the schedule of assessments and exclude the visits that are not needed, as well as all unscheduled visits if you are merging with SV, prior to performing the merge.

```
data adqs_sv;
  merge qs (in = qs) sdtm.sv (in = sv keep = usubjid visit visitnum epoch
                              where = (epoch = 'TREATMENT' and
                                       index(visit, 'UNSCHEDULED' = 0)));
  by usubjid visitnum;
  if sv and not qs then paramcd = 'QSALL';
run;
```

After you have done that, you have created an equivalent of a ‘QSALL’ record manually and now you can just repeat the steps from the previous example to convert this added record into multiple records for different scales. As you can see from the code above, we have set PARAMCD = ‘QSALL’ to make our previous code even more reusable.

Phantom baseline records

One more special case is a baseline visit. While the majority of the visits often retain their collected names as analysis visit names quite often, it is also a common practice to rename whichever visit that contains a baseline flag (ABLFL = ‘Y’) into ‘BASELINE’. However, recall, that a baseline value is the last non-missing value prior to treatment start date, which means that a phantom record could never get a baseline flag as ‘Y’ by design since it’s analysis value is always missing. Take a look at Displays 6 and 7 one more time. The patient in that example did not have the baseline visit, and even after phantom records were added to the dataset there is still no record with ABLFL = ‘Y’ and hence with AVISIT = ‘BASELINE’.

However, we do want to be able to summarize the number of patients who don’t have a baseline assessment in our completion rate table. A solution to this would be to add one more step to our routine and add baseline records for all subjects in the analysis population of interest (let’s assume it to be ITT in our case), since all of those subjects are expected to have at least a baseline assessment. The code below can help us with that:

```
data baseline;
  set adam.adsl (where = (ittfl = 'Y'));
  length paramcd $8. param mcrit2 mcrit2ml $200 avisit $40;
  TRTP   = TRT01P;
  TRTA   = TRT01A;
  paramtyp = 'DERIVED';
  dtype    = 'PHANTOM';
  avisit   = 'BASELINE';
  avisitn  = 1;
  mcrit2   = 'EORTC-QLQ Completion Status';
  do paramcd = 'QLQPFSC', 'QLQRFSC', 'QLQEFSC', 'QLQCFSC', 'QLQSFSC',
               'QLQFASC', 'QLQNVSC', 'QLQPASC', 'QLQQLSC';
    mcrit2ml = 'None of the questions completed';
    param    = put(paramcd, qsparam.);
    output;
  end;
run;
```

```
proc sort data = baseline;
  by usubjid paramcd avisitn;
run;

proc sort data = adqs_ph;
  by usubjid paramcd avisitn;
run;

data adqs_bl;
  merge baseline adqs_ph;
  by usubjid paramcd avisitn;
run;
```

The above mentioned steps would result in ADQS dataset now having baseline records for all subjects in ITT population:

	USUBJID	AVISIT	AVISITN	PARAM	PARAMCD	PARAMTYP	DTYPE	AVAL	ABLFL	MCRIT2ML
26	XXX-001-10102	BASELINE	1	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
27	XXX-001-10102	CYCLE 1 DAY 1	2	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
28	XXX-001-10102	CYCLE 2 DAY 1	3	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
29	XXX-001-10102	CYCLE 3 DAY 1	4	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
30	XXX-001-10102	CYCLE 4 DAY 1	5	Cognitive Functioning Score	QLQCFSC	DERIVED		83		All questions completed
31	XXX-001-10102	END OF TREATMENT	8000	Cognitive Functioning Score	QLQCFSC	DERIVED		50		All questions completed
32	XXX-001-10102	BASELINE	1	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
33	XXX-001-10102	CYCLE 1 DAY 1	2	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
34	XXX-001-10102	CYCLE 2 DAY 1	3	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
35	XXX-001-10102	CYCLE 3 DAY 1	4	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
36	XXX-001-10102	CYCLE 4 DAY 1	5	Emotional Functioning Score	QLQEFSC	DERIVED		75		All questions completed
37	XXX-001-10102	END OF TREATMENT	8000	Emotional Functioning Score	QLQEFSC	DERIVED		75		All questions completed
38	XXX-001-10102	BASELINE	1	Physical Functioning Score	QLQPFSC	DERIVED	PHANTOM	.		None of the questions completed
39	XXX-001-10102	CYCLE 1 DAY 1	2	Physical Functioning Score	QLQPFSC	DERIVED	PHANTOM	.		None of the questions completed
40	XXX-001-10102	CYCLE 2 DAY 1	3	Physical Functioning Score	QLQPFSC	DERIVED	PHANTOM	.		None of the questions completed
41	XXX-001-10102	CYCLE 3 DAY 1	4	Physical Functioning Score	QLQPFSC	DERIVED	PHANTOM	.		None of the questions completed
42	XXX-001-10102	CYCLE 4 DAY 1	5	Physical Functioning Score	QLQPFSC	DERIVED		93		All questions completed
43	XXX-001-10102	END OF TREATMENT	8000	Physical Functioning Score	QLQPFSC	DERIVED		60		All questions completed

Display 8. Example of ADaM.ADQS dataset with a baseline visit added

One thing you can notice, however, is that now this patient has both a baseline record and a Cycle 1 Day 1 one. In case of a regular non-missing assessment, the later would have been converted to a baseline record, since it would have had a baseline flag as 'Y'. In our case though, the Cycle 1 Day 1 record is simply unnecessary, because under normal circumstances patients wouldn't have this visit anymore. Let's assume that according to the study protocol all questionnaires at Cycle 1 Day 1 visit should be filled out before the dosing, meaning that Cycle 1 Day 1 is always expected to be changed to baseline. We cannot influence the existing data and the potential data issues, so it is still possible that some patients will end up having a Cycle 1 Day 1 record that does not qualify as baseline. However, since the phantom rows are imputed by us, we want them to follow the logic that is expected according to the protocol.

The solution is simple: we just remove the Cycle 1 Day 1 record if it's a phantom one by adding one extra line of code to the last data step from the previous example:

```
data adqs_bl;
  merge baseline adqs_ph;
  by usubjid paramcd avisitn;
  if avisitn = 2 and missing(aval) then delete;
run;
```

Finally, the dataset looks as it should:

	USUBJID	AVISIT	AVISITN	PARAM	PARAMCD	PARAMTYP	DTYPE	AVAL	ABLFL	MCRIT2ML
26	XXX-001-10102	BASELINE	1	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
27	XXX-001-10102	CYCLE 2 DAY 1	3	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
28	XXX-001-10102	CYCLE 3 DAY 1	4	Cognitive Functioning Score	QLQCFSC	DERIVED	PHANTOM	.		None of the questions completed
29	XXX-001-10102	CYCLE 4 DAY 1	5	Cognitive Functioning Score	QLQCFSC	DERIVED		83		All questions completed
30	XXX-001-10102	END OF TREATMENT	8000	Cognitive Functioning Score	QLQCFSC	DERIVED		50		All questions completed
31	XXX-001-10102	BASELINE	1	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
32	XXX-001-10102	CYCLE 2 DAY 1	3	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
33	XXX-001-10102	CYCLE 3 DAY 1	4	Emotional Functioning Score	QLQEFSC	DERIVED	PHANTOM	.		None of the questions completed
34	XXX-001-10102	CYCLE 4 DAY 1	5	Emotional Functioning Score	QLQEFSC	DERIVED		75		All questions completed
35	XXX-001-10102	END OF TREATMENT	8000	Emotional Functioning Score	QLQEFSC	DERIVED		75		All questions completed
36	XXX-001-10102	BASELINE	1	Physical Functioning Score	QLQPFSC	DERIVED	PHANTOM	.		None of the questions completed
37	XXX-001-10102	CYCLE 2 DAY 1	3	Physical Functioning Score	QLQPFSC	DERIVED	PHANTOM	.		None of the questions completed
38	XXX-001-10102	CYCLE 3 DAY 1	4	Physical Functioning Score	QLQPFSC	DERIVED	PHANTOM	.		None of the questions completed
39	XXX-001-10102	CYCLE 4 DAY 1	5	Physical Functioning Score	QLQPFSC	DERIVED		93		All questions completed
40	XXX-001-10102	END OF TREATMENT	8000	Physical Functioning Score	QLQPFSC	DERIVED		60		All questions completed

Display 9. Example of ADaM.ADQS dataset after all the steps

CONCLUSION

In conclusion, the addition of phantom records to ADaM datasets can be a valuable technique for addressing data quality issues and improving the accuracy of statistical analyses in clinical trials. Following CDISC standards developed for phantom records ensures that the resulting dataset is transparent, consistent, and easily interpretable. By introducing synthetic observations into the dataset, phantom records imputation technique can help to fill in missing data points and lead to more robust and reliable statistical analyses. We have discussed some common cases where phantom records can be of use, as well as some frequently observed issues with their implementation. However, it is always important to account for your particular scenarios that you encounter in your study, handle them appropriately and make the adjustments whenever necessary. Overall, the use of phantom records can be a powerful tool for dealing with missing data and can ultimately lead to more meaningful insights and better-informed decision-making in clinical trials.

REFERENCES

[1] CDISC ADaM Controlled Terminology, 2022-06-24
https://evs.nci.nih.gov/ftp1/CDISC/ADaM/ADaM%20Terminology.pdf

[2] N. Aaronson, S. Ahmedzai, B. Bergman, et al, The European Organisation for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute 1993; **85**: 365-376.

RECOMMENDED READING

- *CDISC. (2021). Study Data Tabulation Model Implementation Guide for Human Clinical Trials, version 3.4. https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-3-4*
- *CDISC. (2011). Analysis Data Model Implementation Guide, version 1.3. https://www.cdisc.org/standards/foundational/adamig/adamig-1-3*

CONTACT INFORMATION

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

Anastasiia Drach
Intego Group
anastasiia.drach@intego-group.com