PhUSE 2016

Paper CD06

# Analysis Datasets for Baseline Characteristics and Exposure Time for Multi-stage Clinical Trials

Kimberly Umans, Biogen, Cambridge, USA
Katherine Riester, Biogen, Cambridge, USA
Manjit McNeill, Biogen, Maidenhead, UK
Wei Liu, Biogen, Cambridge, USA
Lukasz Kniola, Biogen, Maidenhead, UK

## ABSTRACT

Baseline characteristics and exposure time are important factors in the statistical analysis of clinical trial data; however, their values depend on the reference date of interest. Multi-stage studies or integrated analyses often have multiple, competing reference dates from which to measure these variables. In this situation, the analysis datasets that hold baseline and exposure time data must have a structure that facilitates analyses with variable derivation that is clear to regulatory reviewers. Following the CDISC Analysis Data Model (ADaM), we developed efficient datasets for baseline characteristics (ADBASE) and exposure time (ADTEXP), moving these variables out of the subject-level dataset ADSL. ADBASE has the structure of one record per subject per baseline type, while ADTEXP has a Basic Data Structure (BDS) structure of one record per subject per exposure time parameter. As a result of using these datasets, programming was simplified and greater clarity and transparency was achieved.

## INTRODUCTION AND BACKGROUND

Baseline characteristics (e.g., age at study entry) and exposure time (e.g., time from first to last dose of study treatment) are important factors in the statistical analysis of clinical trial data. In simple studies, these subject-level variables are typically placed in the subject-level ADaM dataset ADSL. However, in a multi-stage trial or for integrated cross-study analyses, baseline characteristics may change from one reference point of interest to another. For example, in a cross-over design, a subject's age at the start of Period 2 of the study may be different from the age at the start of Period 1. Both ages may be needed for data analysis. Similarly, the number of exposure time parameters of interest can be expected to increase as the complexity of the study grows.

In a multi-stage study, retaining all relevant baseline characteristics and exposure time variables in the subject-level dataset may become problematic and result in an unwieldy ADSL. A large number of similarly named variables in ADSL may lead to confusion concerning what reference point was used for the derivation of each variable and which variable to use in a given analysis.

In this paper, we describe ADaM datasets ADBASE and ADTEXP designed to hold baseline characteristics and exposure time variables, respectively. These important variables are thus placed in 'longer, narrower' datasets as opposed to a 'wider' ADSL.

### DISCLAIMER

The scope of this paper is to present the opinions and suggestions of the authors. The interpretations of CDISC standards contained in this paper are those of the authors and are not necessarily correct. Any views and recommendations stated within this paper are those of the authors, and they do not represent the positions of their employer nor the CDISC consortium.

## ILLUSTRATIVE STUDY EXAMPLE

Suppose that a clinical study has two stages, in which Period 1 is randomized, double-blind and placebo-controlled, followed by an open-label single-arm Period 2 in which all subjects receive active drug. Additionally, assume that the study contains an optional substudy, where participants may begin the substudy at any time during Period 1 or Period 2.

Such a study includes four reference dates of interest: start of Period 1, start of Period 2, start of active drug (equaling the start of Period 1 for subjects randomized to active drug in Period 1, or the start of Period 2 for subjects randomized to placebo in Period 1), and finally start of the substudy. Baseline values at each separate reference date may be needed for analysis of study outcomes. Exposure time parameters in such a study may include time on treatment overall, time on study overall, time on treatment in each stage and in the substudy, time on study in each stage and in the substudy, time on study prior to alternative therapy, et cetera.

## ADBASE

Dataset ADBASE has the structure of one record per subject per baseline type. With this structure, it is neither a subject-level dataset nor a parameter-based BDS dataset. Rather, it generally follows the ADaM structure for occurrence data (OCCDS) similar to an occurrence-based dataset such as ADAE. However, ADBASE differs from a typical OCCDS dataset in that all ADBASE records are derived rather than being read in from any SDTM dataset. For this reason the usually required numeric variable xxSEQ (where xx is the pertinent SDTM domain) is not present; instead, numeric variable BASETYPN is used to uniquely identify records within a subject.

Working in the context of the OCCDS structure, a subject is considered to have an "occurrence" of the baseline type if the particular baseline type is applicable to the subject. For example, a subject withdrawing from the study during Period 1 would have an occurrence of Period 1 baseline type but not have the occurrence of Period 2 baseline. If this subject was on active drug in Period 1, they would also have an occurrence and thus an ADBASE record for the active drug baseline type. The subject would have an occurrence of substudy baseline only if the subject participated in the optional substudy.
Key variables for ADBASE include the following:

- Baseline Type (BASETYPE), describing each point of reference,
- Baseline Reference Date (BASERFDT), the reference date for each baseline type.

Each subject has records only for applicable baseline types; as mentioned above, a subject who withdrew prior to the start of Period 2 would not have a record for the Period 2 baseline type. Sample baseline characteristics could include age at baseline, disease severity, past history of co-morbidities, weight, or past use of concomitant medication. Each characteristic is derived using BASERFDT as the reference date, as illustrated in Table 1.

**Table 1: SAMPLE ADBASE SPECIFICATION**

| VARIABLE | LABEL | CODELIST | DERIVATION |
|---|---|---|---|
| *%idvar* | | | *ADSL.*<br>*Standard subject identification variables from ADSL would appear first* |
| BASETYPE | Baseline Type | "Period 1", "Period 2", "Start of Active Drug", "Substudy" | Create one record for each baseline type appropriate to the subject |
| BASETYPN | Baseline Type (N) | 1, 2, 3, 4 | In correspondence with BASETYPE |
| BASERFDT | Baseline Reference Date | | Date of first dose of treatment for the designated BASETYPE |
| COUNTRY | Country | | DM.COUNTRY or ADSL.COUNTRY |
| AAGE | Analysis Age | | Age at BASERFDT |
| WEIGHTBL | Baseline Weight | | Value of last weight measurement on or prior to BASERFDT |
| VASBL | Baseline VAS Score | | Last visual analogue scale score on or prior to BASERFDT |
| DIAGMOS | Months Since Diagnosis | | Number of months from first diagnosis to BASERFDT |
| HISTHDFL | History of Headaches Flag | Y, N | Y if subject has medical history of headaches, or a headache adverse event with start date prior to BASERFDT |
| PMEDC1FL | Prior Medication Category 1 Flag | Y, N | Y if subject has relevant treatment history in medication category 1, or concomitant medication in medication category 1 with start date prior to BASERFDT |

Note that in this sample specification, a variable for country was included even though its value is not expected to change over the course of the study or across baseline types. Whether to include such variables in ADBASE, to keep them in ADSL, or include in both datasets would be a matter of preference. Including time-constant baseline variables in ADBASE has the advantage of allowing for a more self-contained ADBASE dataset. Obviously, all required ADSL variables must be retained in ADSL.

A brief example of an ADBASE dataset is presented in Table 2. In this invented example, subject 001 was randomized to active drug in Period 1 and participated in both study periods. Subject 002 withdrew during Period 1, and subject 003 was randomized to placebo in Period 1 and enrolled in the substudy during Period 2.

**Table 2: SAMPLE ADBASE**

| SUBJID | BASETYPN | BASETYPE | BASERFDT | AAGE | WEIGHTBL | VASBL | DIAGMOS | etc. |
|---|---|---|---|---|---|---|---|---|
| 001 | 1 | Period 1 | 01Sep2014 | 24 | 70 | 64 | 15 | |
| 001 | 2 | Period 2 | 02Jan2015 | 25 | 71 | 60 | 19 | |
| 001 | 3 | Start of Active Drug | 01Sep2014 | 24 | 70 | 64 | 15 | |
| 002 | 1 | Period 1 | 11Oct2014 | 50 | 67 | 74 | 31 | |
| 003 | 1 | Period 1 | 01Nov2014 | 48 | 82 | 59 | 20 | |
| 003 | 2 | Period 2 | 01Mar2015 | 48 | 81 | 63 | 24 | |
| 003 | 3 | Start of Active Drug | 01Mar2015 | 48 | 81 | 63 | 24 | |
| 003 | 4 | Substudy | 28Mar2015 | 49 | 81 | 67 | 25 | |

**USE OF ADBASE**

Once ADBASE is restricted to a particular baseline type, the subset dataset has only one record per subject and can be merged easily with ADSL and ADaM BDS datasets, enabling use of ADBASE variables as covariates in statistical models.

Moreover, in the context of a multi-stage study, baseline type variables such as BASETYPE may appear in other ADaM datasets to specify which baseline was used to derive values. For example, BASETYPE may be included in ADTTE where time to event is calculated based on different reference dates. Similarly, BASETYPE may be included in ADAE where treatment-emergence and relative start day is based on different baseline dates.

In these other ADaM datasets with variable BASETYPE, as with ADBASE, each subject has records only for baseline types applicable to the subject. Merging ADBASE with another dataset that contains BASETYPE, on subject and baseline type, adds the correct baseline characteristics to the dataset for analyses.

## ADTEXP

Standard ADSL variables TRxxDURD, TRxxDURW, etc., are available to represent time on treatment by sequential study period. However, several other measures of exposure may also be needed for study analyses. For example, it may be of interest to analyze both time taking treatment (first to last dose) and time exposed to treatment (first to last dose plus some amount of time based on PK/PD effects). Additional analyses may require participation time to be censored at certain event occurrences such as disease recurrence or receipt of alternate therapy. When the number of exposure or participation time parameters exceeds a reasonable number for inclusion in ADSL, dataset ADTEXP is proposed to store these duration measures.

Dataset ADTEXP has the standard BDS structure of one record per subject per parameter. Key variables include

- Parameter (PARAM),
- Analysis Start Date (ASTDT) and Analysis End Date (AENDT),
- Analysis Value (AVAL).

In addition to the dataset specification, necessary metadata for ADTEXP includes a parameter lookup table which designates the start and stop dates and units for each parameter. A sample dataset specification is displayed in Table 3 together with a sample lookup table in Table 4.

**Table 3: SAMPLE ADTEXP SPECIFICATION**

| VARIABLE | LABEL | CODELIST | DERIVATION |
|---|---|---|---|
| *%idvar* | | | *ADSL.*<br>*Standard subject identification variables from ADSL would appear first* |
| PARAM | Parameter | see parameter lookup table | see parameter lookup table |
| PARAMN | Parameter (N) | see parameter lookup table | see parameter lookup table |
| PARAMCD | Parameter Code | see parameter lookup table | see parameter lookup table |
| ASTDT | Analysis Start Date | | see parameter lookup table |
| AENDT | Analysis End Date | | see parameter lookup table |
| AVAL | Analysis Value | | AENDT minus ASTDT + 1 day, converted to appropriate units per parameter lookup table. |

**Table 4: SAMPLE ADTEXP PARAMETER LOOKUP TABLE**

| PARAMN | PARAMCD | PARAM | units | ASTDT | AENDT |
|---|---|---|---|---|---|
| 1 | ST01DURD | Time on Study Period 1 (days) | days | ADSL.AP01SDT | ADSL.AP01EDT |
| 2 | ST01DURW | Time on Study Period 1 (weeks) | weeks | ADSL.AP01SDT | ADSL.AP01EDT |
| 3 | ST02DURD | Time on Study Period 2 (days) | days | ADSL.AP02SDT | ADSL.AP02EDT |
| 4 | STACDURD | Time on Study Active (days) | days | ADSL.AP01SDT or ADSL.AP02SDT | ADSL.AP02EDT |
| 5 | STS1DURD | Time on Study Substudy 1 (days) | days | ADSL.SS01SDT | ADSL.SS01EDT |
| 6 | TR01DURD | Time on Treatment Period 1 (days) | days | ADSL.TR01SDT | ADSL.TR01EDT |
| 7 | TE01DURD | Time Exposed to Treatment Period 1 (days) | days | ADSL.TR01SDT | ADSL.TR01EDT + xx days, or ADSL.AP01EDT whichever first |
| 8 | TP01DURD | Time on Treatment Period 1 Prior to Alternative Therapy (days) | days | ADSL.TR01SDT | Date of first alternative therapy or ADSL.AP01EDT, whichever first |

Because the meaning of the exposure or participation time measure is stored in the PARAM variable (up to 200 characters), a more detailed description is possible in the ADTEXP dataset compared to using a 40 character maximum variable label in ADSL. This advantage becomes particularly important as the complexity and quantity of similarly-derived exposure time measures needed for analysis exceeds a small number.

All start and end dates needed for ASTDT and AENDT are considered key study dates and thus expected to be contained in ADSL. In this case, ADSL is the only input dataset needed to create ADTEXP. A record is created for a subject and parameter provided the start and end dates for the parameter are not missing for the subject. In the fictional example in Table 5, subject 001 participated in both study periods but not the substudy and did not receive alternative therapy.

**Table 5: SAMPLE ADTEXP**

| SUBJID | PARAMN | PARAM | ASTDT | AENDT | AVAL |
|---|---|---|---|---|---|
| 001 | 1 | Time on Study Period 1 (days) | 01Sep2014 | 02Jan2015 | 124 |
| 001 | 2 | Time on Study Period 1 (weeks) | 01Sep2014 | 02Jan2015 | 17.714 |
| 001 | 3 | Time on Study Period 2 (days) | 02Jan2015 | 01May2015 | 120 |
| 001 | 4 | Time on Study Active (days) | 01Sep2014 | 01May2015 | 243 |
| 001 | 6 | Time on Treatment Period 1 (days) | 02Sep2014 | 27Dec2014 | 117 |
| 001 | 7 | Time Exposed to Treatment Period 1 (days) | 02Sep2014 | 02Jan2015 | 123 |

## CONCLUSION

A fundamental principle of the Analysis Data Model is that the structure and content of the analysis datasets must support clear, unambiguous communication of the scientific and statistical aspects of the trial. Multi-stage studies, or integrated analyses across clinical trials, have multiple reference points for baselines of interest and numerous measures of duration of subject participation needed for statistical analysis of the clinical data. Consequently, the number of variables required to represent baseline characteristics and exposure time measures may increase to the point where housing such variables in ADSL is impractical, causing ADSL to become unwieldy and thus raising the risk of potential errors.

Datasets ADBASE and ADTEXP are proposed as possible alternative repositories, respectively, for baseline characteristics that change over study reference dates and for participation duration measures. When the complexity of the clinical trial warrants their use, greater clarity and transparency can be achieved.

## REFERENCES


- Analysis Data Model version 2.1; Clinical Data Interchange Standards Consortium, published 2009-12-17
- Analysis Data Model Implementation Guide version 1.1; Clinical Data Interchange Standards Consortium, published 2016-02-12
- ADaM Structure for Occurrence Data (OCCDS) version 1.0; Clinical Data Interchange Standards Consortium, published 2016-02-12


## CONTACT INFORMATION

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


Kimberly Umans / Katherine Riester / Wei Liu
Biogen
250 Binney Street
Cambridge MA 02453
USA
Email: kim.umans@biogen.com; katherine.riester@biogen.com; wei.liu@biogen.com

Lukasz Kniola / Manjit McNeill
Biogen Idec Ltd
Innovation House
70 Norden Road
Maidenhead, Berkshire SL6 4AY
UK
Email: lukasz.kniola@biogen.com; manjit.mcneill@biogen.com