PhUSE US Connect 2018

Paper SD16

ADaM Intermediate Dataset: how to improve your analysis traceability

Angelo Tinazzi, Cytel Inc., Geneva, Switzerland
Teresa Curto, Cytel Inc., Waltham, USA
Ashish Aggarwal, Cytel Inc. Pune, India

ABSTRACT

An important component of a regulatory review is an understanding of the provenance of the data. “Traceability permits an understanding of the relationships between the analysis results, analysis datasets, tabulation datasets, and source data” (FDA SDTCG October 2017).
Very often the derivation of your endpoint requires several ‘complex’ steps where you need to make sure all the steps are traceable in the analysis datasets (ADaM) and in the documentation you will create. A common recommended approach used in some of the available CDISC TAUGs is the use of an “Intermediate” ADaM dataset e.g. for the derivation of a TTE endpoint.
The aim of this presentation is to illustrate a recent analysis where a complex endpoint derivation required the use of intermediate ADaM datasets and how the steps involved in the creation of such intermediate ADaM datasets were documented in the Analysis Data Reviewer Guide (ADRG).

INTRODUCTION

Several traceability definitions exist as reported in Table 1.

Source	Definition
Wikipedia	Traceability is the ability to verify the history, location, or application of an item by means of documented recorded identification Ref. "Glossary," *ASME Boiler and Pressure Vessel Code*, Section III, Article NCA-9000
GCP	All clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation and verification Ref. ICH E6 GCP Section 2.10 “GCP Principles”
CDISC	The property that enables the understanding of the data’s lineage and/or the relationship between an element and its predecessor(s). Ref. CDISC ADaM Ig 1.1
FDA	An important component of a regulatory review is an understanding of the provenance of the data (i.e., traceability of the sponsor’s results back to the CRF data). Traceability permits an understanding of the relationships between the analysis results, analysis datasets, tabulation datasets, and source data. Ref. FDA Study Data Technical Conformance Guide (SDTCG)

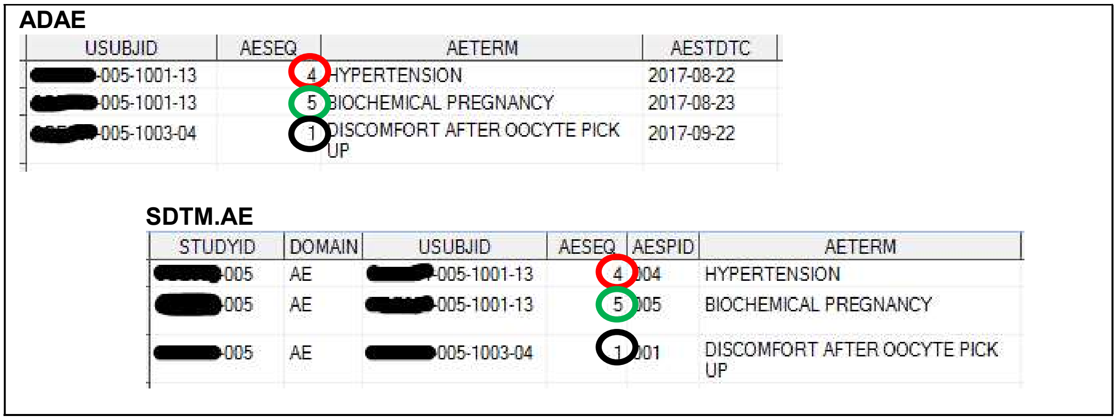
Table 1: Traceability Definitions

As you can see this is not a new topic and actually it is the “leitmotiv” of everything we do in our daily work starting from the data acquisition step (as mentioned in the GCP), till the publication of our results e.g. in a Clinical Study Report.
The FDA has re-enforced the importance of traceability in the SDTCG [1] “*Based upon reviewer experience, establishing traceability is one of the most problematic issues associated with any data conversion. If the reviewer is unable to trace study data from the data collection of subjects participating in a study to the analysis of the overall study data, then the regulatory review of a submission may be compromised*” and also an FDA reviewer in a public webinar affirmed that “*Traceability may be even more important than Data Standards (To me)*” [2].

HOW TO ACHIEVE TRACEABILITY

Traceability can be achieved through two main methods:

- data-point traceability: p*oints directly to the specific predecessor record(s) and should be implemented if practical and feasible*. For example when creating an ADAM dataset for AE, e.g. ADAE, this will probably be derived from the SDTM AE dataset and it will be "worth it" to copy over from SDTM to ADaM the AESEQ variable pointing to the source record in SDTM AE dataset (see example in figure 1).



ADAE

USUBJID	AESEQ	AETERM	AESTDTC
[redacted]-005-1001-13	4	HYPERTENSION	2017-08-22
[redacted]-005-1001-13	5	BIOCHEMICAL PREGNANCY	2017-08-23
[redacted]-005-1003-04	1	DISCOMFORT AFTER OOCYTE PICK UP	2017-09-22

SDTM.AE

STUDYID	DOMAIN	USUBJID	AESEQ	AESPID	AETERM
[redacted]-005	AE	[redacted]-005-1001-13	4	004	HYPERTENSION
[redacted]-005	AE	[redacted]-005-1001-13	5	005	BIOCHEMICAL PREGNANCY
[redacted]-005	AE	[redacted]005-1003-04	1	001	DISCOMFORT AFTER OOCYTE PICK UP

Figure 1: Traceability between ADAE and SDTM.AE

- metadata traceability: *facilitates the understanding of the relationship of the analysis variable to its source dataset(s) and variable(s)*. This traceability is established by describing (via metadata) the algorithm used or steps taken to derive or populate an analysis variable from its immediate predecessor. Metadata traceability is also used to establish the relationship between an analysis result and ADaM dataset(s). Metadata traceability should be applied when data-point traceability is not feasible or in any in case in support of data-point traceability e.g. in the define.xml or in the ADRG when complex derivations are applied [3].

Table 2 summarizes the current available 'methods' in CDISC to support traceability.

Standard	Data Point Traceability	Metadata Traceability (Supportive Documents)
SDTM	«Not applicable/Not needed»	aCRF define.xml SDRG
ADaM	Copy SDTM variables --SEQ from SDTM SRCDOM/SRCVAR/SRCSEQ ADTF ASEQ DTYPE ANLxxFL Occurrence Flags in OCCDS Intermediate ADaMs	define.xml ADRG SAP
Analysis Results	«Not applicable/Not needed»	define.xml (ARM extension) ADRG SAP

Table 2: How to be Traceable

BE ANALYSIS-READY

Among the four ADaM fundamental principles, having analysis datasets analysis-ready, and as much as possible one-proc-away, is also a way of improving traceability; this principle requires that *at minimum the analysis datasets contain the data needed for the review and re-creation of specific statistical analyses*. As such the analysis datasets should contain all the records and variables needed by the analysis. For example consider the following code:

```
data Age_Table;
  set adam.adsl;
  age=int((trt02sdt-input(brthdtc,yymmdd10.))/ 365.25);
run;
proc means data=Age_Table;
  class trt01p;
  var age;
run;
```

The analysis dataset used in this code, ADSL, can be not considered analysis-ready. The hypothetical SAP of this study is requesting a demographics analysis where age is re-calculated from the treatment start date of the second treatment period (e.g. trt02sdt) as opposed to the AGE variable available in ADSL and copied from SDTM.DM that was collected or derived at study entry e.g. at the time of informed consent. The analysis dataset in this case to be considered analysis-ready should have a specific derived age variable already available in ADSL, for example AAGE and not "embedding" the derivation in the analysis table program.

MAKE USE OF INTERMEDIATE ANALYSIS DATASET

With the ADaM Implementation Guidance 1.1 [4], but also in some Therapeutic Area Guidance (TAUG) [5][6], the use of "Intermediate" analysis datasets have been recommended to improve the data flow understanding when complex data transformation are needed in support of a statistical analysis. From ADaM Ig 1.1:

> *"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. "*

This is moreover important when we need to handle several steps e.g. a complex algorithm that could also increase / multiply the number of records the ADaM dataset could have especially if we want to guarantee a good level of traceability, thus keeping original records in our ADaM datasets (e.g. original SDTM records).

Intermediate Analysis Dataset Example: Derivation of a Composite Time-to-Event (TTE) Endpoint

A common situation in which an Intermediate ADaM dataset might be needed is with TTE endpoint derivation, where there is the need to verify that the correct date was selected (e.g. with a composite TTE endpoint); in this situation it can be "advantageous" to have all "candidate" dates available for all subjects in one Intermediate analysis dataset prior to creating the TTE variables in the final TTE ADaM dataset.

The example in figure 2 and 3 is an 'elaborated' version taken from the Breast TAUG [6] (see in particular section 5.3.3). In there they wanted to analyse the primary endpoint 'Progression Free Survival (PFS)', a time-to-event endpoint where the event of interest is either the progression of the tumour disease or the death, whichever came first. In this endpoint we might want to censor (not consider) events occurred after the administration of some kind of 'prohibited' medications (follow-up anti-cancer therapies). The censored events are censored at the date of last evaluable tumour progression. There are several complexity factors in this definition:

- several dates to be considered in the derivation of the endpoint
- not all potential dates have to be considered
- among multiple events only the first occurrence needs to be selected

Therefore we need to find a way of making this date/event 'selection' process as transparent as possible so that the reviewer can clearly understand which event triggered the TTE endpoint, which date was used to calculate the time part of the TTE endpoint and in case an event was not considered as such make clear the rationale. All this can be achieved through metadata (e.g. define.xml plus analysis reviewer guide), but it would be also good if this can be traceable with data-point too. Moreover, this later will make the process easier for review and reproduce the results (Quality Control process).

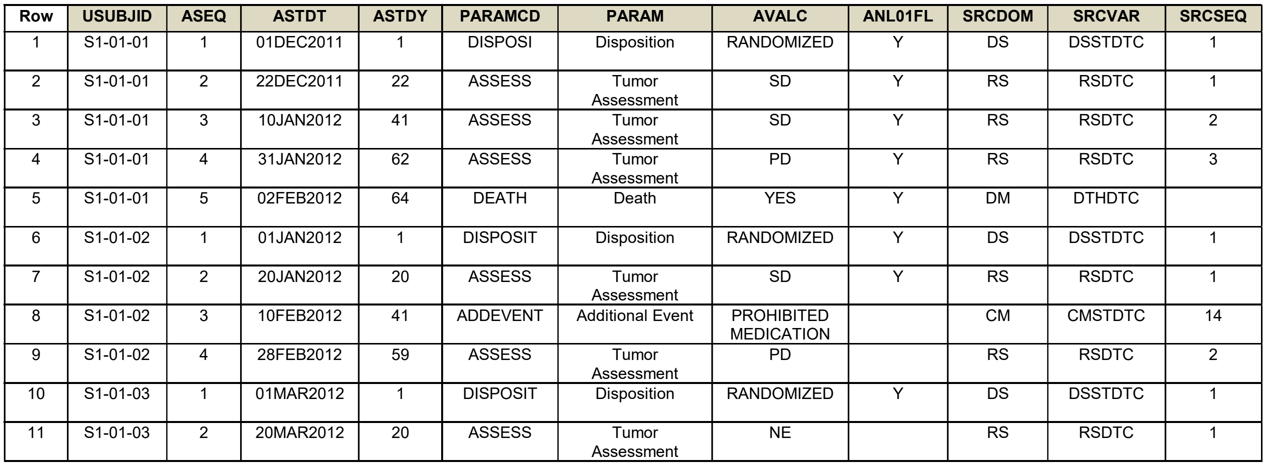
In order to achieve this aimed full transparency (full traceability), an intermediate dataset where all potential dates are stored, is proposed.

Figure 2 shows an example of an intermediate dataset named ADTTEDAT where we store all dates from SDTM that could potentially contribute to the TTE endpoint, including the date of randomization from the disposition dataset that will determine the starting point of our time-to-event endpoint (e.g. time=(date of event/censor-randomization date)+1). The "candidate" dates are as follows:

- date of randomization (start of the TTE endpoint) (PARAMCD=DISPOSIT and AVALC=RANDOMIZED).
- each individual tumor assessment (PARAMCD=ASSESS).
- any prohibited medications (PARAMCD=ADDVENT and AVALC= PROHIBITED MEDICATION).
- date of death (PARAMCD=DEATH).
- All these dates have to be included in the ADTTEDAT dataset but only those occurred prior to the

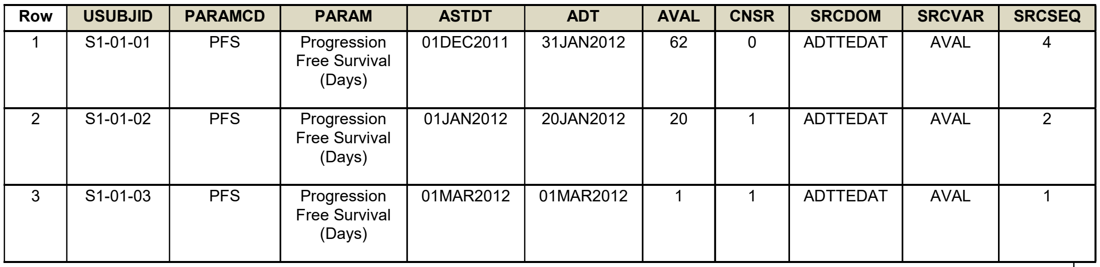
prohibited medication intake have to be considered in the derivation of the PFS TTE endpoint. The ‘eligible’ dates are flagged with an analysis flag (ANL01FL=Y).

In this dataset we also kept the three SRC--- variables to be able to identify the source SDTM record.



Row	USUBJID	ASEQ	ASTDT	ASTDY	PARAMCD	PARAM	AVALC	ANL01FL	SRCDOM	SRCVAR	SRCSEQ
1	S1-01-01	1	01DEC2011	1	DISPOSI	Disposition	RANDOMIZED	Y	DS	DSSTDTC	1
2	S1-01-01	2	22DEC2011	22	ASSESS	Tumor Assessment	SD	Y	RS	RSDTC	1
3	S1-01-01	3	10JAN2012	41	ASSESS	Tumor Assessment	SD	Y	RS	RSDTC	2
4	S1-01-01	4	31JAN2012	62	ASSESS	Tumor Assessment	PD	Y	RS	RSDTC	3
5	S1-01-01	5	02FEB2012	64	DEATH	Death	YES	Y	DM	DTHDTC	
6	S1-01-02	1	01JAN2012	1	DISPOSIT	Disposition	RANDOMIZED	Y	DS	DSSTDTC	1
7	S1-01-02	2	20JAN2012	20	ASSESS	Tumor Assessment	SD	Y	RS	RSDTC	1
8	S1-01-02	3	10FEB2012	41	ADDEVENT	Additional Event	PROHIBITED MEDICATION		CM	CMSTDTC	14
9	S1-01-02	4	28FEB2012	59	ASSESS	Tumor Assessment	PD		RS	RSDTC	2
10	S1-01-03	1	01MAR2012	1	DISPOSIT	Disposition	RANDOMIZED	Y	DS	DSSTDTC	1
11	S1-01-03	2	20MAR2012	20	ASSESS	Tumor Assessment	NE		RS	RSDTC	1

Figure 2: ADTTEDAT Intermediate Analysis Dataset



Row	USUBJID	PARAMCD	PARAM	ASTDT	ADT	AVAL	CNSR	SRCDOM	SRCVAR	SRCSEQ
1	S1-01-01	PFS	Progression Free Survival (Days)	01DEC2011	31JAN2012	62	0	ADTTEDAT	AVAL	4
2	S1-01-02	PFS	Progression Free Survival (Days)	01JAN2012	20JAN2012	20	1	ADTTEDAT	AVAL	2
3	S1-01-03	PFS	Progression Free Survival (Days)	01MAR2012	01MAR2012	1	1	ADTTEDAT	AVAL	1

Figure 3: ADTTEPFS Progression Free Survival Analysis Dataset

Figure 3 instead shows the ADTTEPFS TTE analysis dataset that was derived from the intermediate analysis dataset ADTTEDAT, where ADT, the event/censor date, is derived as the date of either first progression occurrence (AVALC=PD where PARAMCD=ASSESS) or death (AVALC=YES where PARAMCD=DEATH), whichever came first, or last valid tumor assessment (AVALC≠NE where PARAMCD=ASSESS); in the first case CNSR=0 indicates the event occurred, while in the second case CNSR=1 indicates the event did not occur; assessments occurred after the intake of prohibited medications (ANL01FL=Null) were not considered, thus the PFS TTE endpoint was censored at the last valid tumor assessment.

For example from figure 2:

- subject S1-01-01 had a PFS event “caused” by the progression of the tumor (PARAMCD=ASSESS and AVALC=PD, see row nr. 4).
- subject S1-01-02 had a progression of the tumor (row nr. 9) but the progression occurred after the intake of the prohibited medication (ANL01FL=Null for row nr 9 because previously, row nr. 8, the patient took the prohibited medication).
- subject S1-01-03 had only one tumor assessment (row nr. 11) after the randomization, but this assessment was judged as ‘NE’ (Not evaluable) and therefore the PFS TTE endpoint was censored at the randomization date since no other date/assessments were available.

This is a perfect example of full traceability, where the reviewer can clearly see which records (and dates) were taken into consideration in the complex derivation of the TTE endpoint and which records were not taken into consideration because not satisfying the endpoint definition.

Table 3 shows other examples of intermediate analysis datasets.

ADaM	Derived from Intermediate ADaM	Comment
ADTTEAE	ADAE	ADTTEAE contains the time to event endpoint related to AE. Because ADAE has been already developed to support standard AE tables (AE incidence for example), ADTTEAE can use ADAE as source. As such ADAE can be considered an intermediate ADaM dataset of ADTTEAE
ADEXSUM	ADEX	For example in oncology one might derive ADEX being a copy of SDTM.EX, where information such as 'cycle' duration, dose in mg/m^2 (if original CRF was collecting the dose in 'mg' only), dose reduction, dose delay, etc. This information is derived only for the purpose of then making some aggregation by patient, by drug if the study 'drug' was made of more than one concomitant investigational drug. The ADEXSUM can be then created from ADEX. ADEXSUM will have a BDS structure where each parameter represents derived (summarized) information per patient, for example total dose received, nr. of cycles received, nr. of dose reductions, nr. of dose delay, and so on. ADEX is then an intermediate ADaM dataset of ADEXSUM and it is not necessarily used as source of any statistical output.

Table 3: Other examples of intermediate analysis datasets

APPLYING THE ADAM INTERMEDIATE DATASETS: A USE CASE

We would like now to present a real case of a study we recently analyzed for one of our sponsor where we have been in a "special" situation where the use of Intermediate Analysis Datasets was needed.

The primary objective of the study was to assess the effects Treatment A vs. Treatment B on blood glucose levels. Blood glucose was collected continually from patients for 3 weeks using a continuous glucose monitor (CGM) device. Measurements were taken in 5 minute increments from 1 week prior to treatment to 2 weeks post-treatment.

AVERAGE BLOOD GLUCOSE CHANGE FROM BASELINE

Average blood glucose was calculated for several time periods and compared to baseline. The following steps were taken, all on a patient level:

1. Each measurement was assigned to a relative hour. For example, relative Hour 1 was from the time of treatment to the next nominal hour. Hour 2 was the next full hour, and so on. This was done for Hour -72 to Hour 360. See Table 4 for an example, where "dose time" is the time of treatment.

Dose Time	Relative Hour	Start Time	End Time
	-4	10:00:01	11:00:00
	-3	11:00:01	12:00:00
	-2	12:00:01	13:00:00
	-1	13:00:01	13:14:59
13:15	1	13:15:00	14:00:00
	2	14:00:01	15:00:00
	3	15:00:01	16:00:00
	4	16:00:01	17:00:00

Table 4: Relative Hour Derivation Example

2. The mean blood glucose for each hour was calculated.
3. Each hour was assigned to a time period using visit windows (see Table 5). The average blood glucose for each time period was calculated by taking the average of the hourly means. The following time periods were derived: Baseline (Hour -72 to hour -1), Day 1 to Day 15 and Day 1-3 combined.
4. The change from baseline was calculated for each time period.

Visit	Start	End
Baseline	Hour -72	Hour -1
Day 1	Hour 1	Hour 24
Day 2	Hour 25	Hour 48
Day 3	Hour 49	Hour 72
Day 4	Hour 73	Hour 96
Day 5	Hour 97	Hour 120
Day 6	Hour 121	Hour 144
Day 7	Hour 145	Hour 168
Day 8	Hour 169	Hour 192
Day 9	Hour 193	Hour 216
Day 10	Hour 217	Hour 240
Day 11	Hour 241	Hour 264
Day 12	Hour 265	Hour 288
Day 13	Hour 289	Hour 312
Day 14	Hour 313	Hour 336
Day 15	Hour 337	Hour 360
Day 1-3	Hour 1	Hour 72

Table 5 Visit Windows

ENDPOINTS

The primary statistical endpoint was the change in average blood glucose from baseline (Hour -72 to Hour 1) to Day 1-3 (Hour 1-72). The analysis of this endpoint was completed using a linear model (ANCOVA) with fixed effects for treatment group. Model covariates were study site and baseline (72-hour) blood glucose average.

Several other secondary and exploratory endpoints were also derived and used for analysis:

1. Area Under the Effect (AUE) for average blood glucose over time - The AUE curve for blood glucose was calculated for the following time points: Hour -72 to Hour 72, Hour -1 to 168 and Hour -1 to 360.
2. Blood glucose categories - The percent of the time the hourly average blood glucose was in five distinct categories. This was derived for each day, from Day 1 to Day 15.
3. Glycemic variability - The coefficient of variation (CV) of the hourly averages within each time period was calculated. First, the % CV for each subject was calculated using the formula (SD/mean)*100. The values from each hourly average glucose measurement over the time period were used. This was derived for Hour 1-24, Hour 1-48, Hour 1-72, Hour 1-168 and Hour 1-360.
4. Maximum hourly blood glucose value – For each time period, the maximum hourly blood glucose was identified. This was based on hourly averages (rather than each individual measurement). This was derived for Hour 1-24, Hour 1-48, Hour 1-72, Hour 1-168 and Hour 1-360.
5. Distance travelled – For each time period, the distance traveled was calculated using this formula: Distance travelled = maximum hourly blood glucose value - 72-hour baseline blood glucose value. This was derived for Hour 1-24, Hour 1-48, Hour 1-72, Hour 1-168 and Hour 1-360.
6. Time to increased and maximum glucose levels – Time to event variables were created for each patient for increased glucose level and maximum glucose level. These were derived for Hour 1-72 and Hour 1-360.
7. Time to the earliest significantly increased blood glucose level in hours was defined as time from administration of study treatment to the first blood glucose level that is ≥ 2 SDs above the 72-hour baseline average blood glucose level (relative hour of first increased blood glucose level). Patients without a ≥ 2 SD increase during the time period will be censored at the end of the period.
8. Time the maximum blood glucose level in hours was defined as the time from administration of study treatment to the maximum hourly average blood glucose level (relative hour of maximum blood glucose level). There was no censoring, as every subject had a maximum value.
9. Cumulative distribution of blood glucose.
10. Maximum blood glucose levels – Maximum hourly average blood glucose levels were also derived for each study day (Day 1 to Day 15).Change from baseline was calculated for each of study day. This endpoint was analyzed in the same way as the primary endpoint.

ADAM IMPLEMENTATION
The chart in figure 4 is extracted from the Analysis Data Reviewer Guide (ADRG) produced in support of the analysis datasets documentation (define.xml).

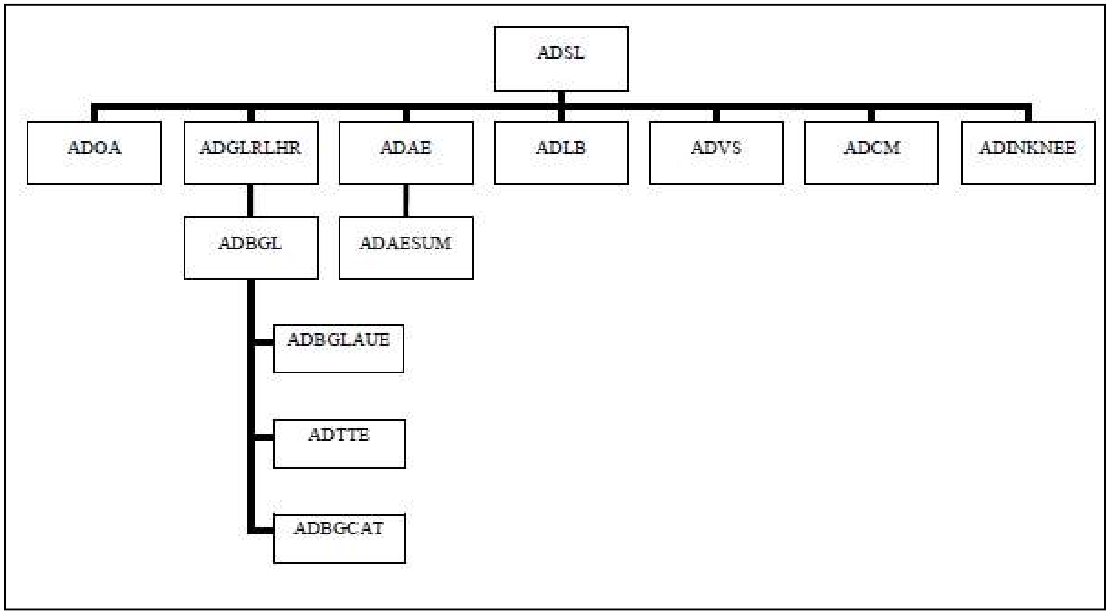
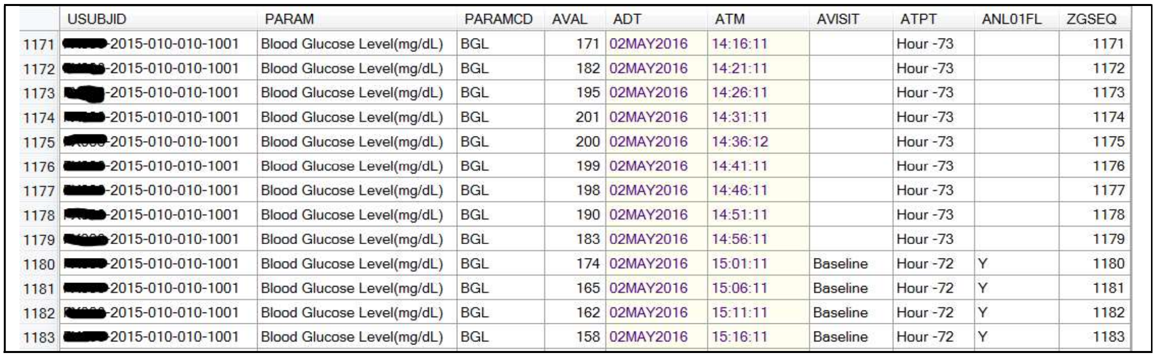


Figure 4: ADaM Data Dependencies

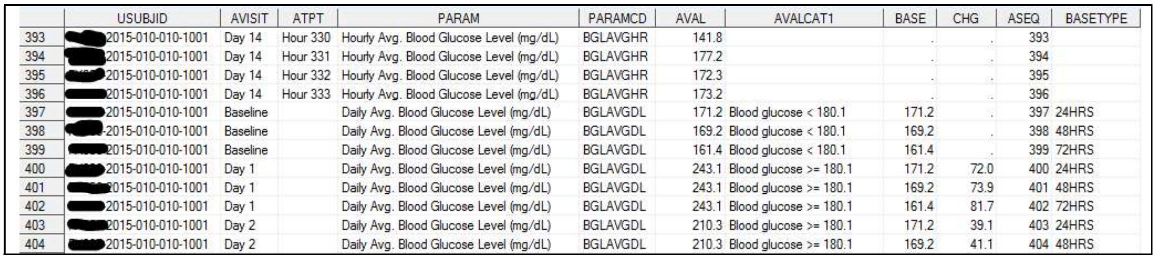
The chart illustrates the data dependency in the ADaM datasets developed in support of the analysis:

- The ADGLRLHR ADaM dataset is the main intermediate dataset created from a sponsor SDTM.ZG dataset we did create in support of all Glucose endpoints; in ADGLRLHR we did calculate the *relative hour*, as per table 4 above, and the *daily means of the Blood Glucose Level* by applying windowing criteria, as per table 5 above [see bullet point 1 to 3 described in the "Study Background" section]. See figure 5 for an example of ADGLRLHR. Of note see the use of ANL01FL to flag those observations that were not used in the analysis (ANL01FL=null) being the observations occurred prior to Baseline / Hour -72.
- ADBGL was then created to support the change from baseline analysis (see figure 6 and bullet point 4 described in the "Study Background, Average Glucose Change from Baseline" section). Of note the addition of the variable ASEQ because the dataset will be then used by other ADaM dataset (see ADTTE described later on); also the BASETYPE variable and the duplication of records to support multiple baseline definitions.



	USUBJID	PARAM	PARAMCD	AVAL	ADT	ATM	AVISIT	ATPT	ANL01FL	ZGSEQ
1171	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	171	02MAY2016	14:16:11		Hour -73		1171
1172	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	182	02MAY2016	14:21:11		Hour -73		1172
1173	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	195	02MAY2016	14:26:11		Hour -73		1173
1174	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	201	02MAY2016	14:31:11		Hour -73		1174
1175	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	200	02MAY2016	14:36:12		Hour -73		1175
1176	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	199	02MAY2016	14:41:11		Hour -73		1176
1177	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	198	02MAY2016	14:46:11		Hour -73		1177
1178	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	190	02MAY2016	14:51:11		Hour -73		1178
1179	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	183	02MAY2016	14:56:11		Hour -73		1179
1180	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	174	02MAY2016	15:01:11	Baseline	Hour -72	Y	1180
1181	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	165	02MAY2016	15:06:11	Baseline	Hour -72	Y	1181
1182	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	162	02MAY2016	15:11:11	Baseline	Hour -72	Y	1182
1183	[redacted]-2015-010-010-1001	Blood Glucose Level(mg/dL)	BGL	158	02MAY2016	15:16:11	Baseline	Hour -72	Y	1183

Figure 5: ADGLRLHR with AVISIT (visit time-point) and ATPT (relative-hour time-point)

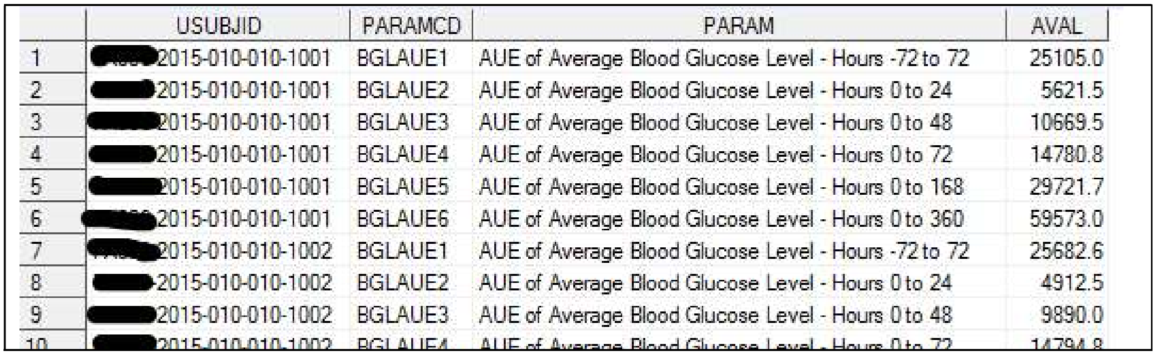


	USUBJID	AVISIT	ATPT	PARAM	PARAMCD	AVAL	AVALCAT1	BASE	CHG	ASEQ	BASETYPE
393	[redacted]2015-010-010-1001	Day 14	Hour 330	Hourly Avg. Blood Glucose Level (mg/dL)	BGLAVGHR	141.8		.	.	393	
394	[redacted]2015-010-010-1001	Day 14	Hour 331	Hourly Avg. Blood Glucose Level (mg/dL)	BGLAVGHR	177.2		.	.	394	
395	[redacted]2015-010-010-1001	Day 14	Hour 332	Hourly Avg. Blood Glucose Level (mg/dL)	BGLAVGHR	172.3		.	.	395	
396	[redacted]2015-010-010-1001	Day 14	Hour 333	Hourly Avg. Blood Glucose Level (mg/dL)	BGLAVGHR	173.2		.	.	396	
397	[redacted]2015-010-010-1001	Baseline		Daily Avg. Blood Glucose Level (mg/dL)	BGLAVGDL	171.2	Blood glucose < 180.1	171.2	.	397	24HRS
398	[redacted]-2015-010-010-1001	Baseline		Daily Avg. Blood Glucose Level (mg/dL)	BGLAVGDL	169.2	Blood glucose < 180.1	169.2	.	398	48HRS
399	[redacted]2015-010-010-1001	Baseline		Daily Avg. Blood Glucose Level (mg/dL)	BGLAVGDL	161.4	Blood glucose < 180.1	161.4	.	399	72HRS
400	[redacted]2015-010-010-1001	Day 1		Daily Avg. Blood Glucose Level (mg/dL)	BGLAVGDL	243.1	Blood glucose >= 180.1	171.2	72.0	400	24HRS
401	[redacted]2015-010-010-1001	Day 1		Daily Avg. Blood Glucose Level (mg/dL)	BGLAVGDL	243.1	Blood glucose >= 180.1	169.2	73.9	401	48HRS
402	[redacted]2015-010-010-1001	Day 1		Daily Avg. Blood Glucose Level (mg/dL)	BGLAVGDL	243.1	Blood glucose >= 180.1	161.4	81.7	402	72HRS
403	[redacted]2015-010-010-1001	Day 2		Daily Avg. Blood Glucose Level (mg/dL)	BGLAVGDL	210.3	Blood glucose >= 180.1	171.2	39.1	403	24HRS
404	[redacted]2015-010-010-1001	Day 2		Daily Avg. Blood Glucose Level (mg/dL)	BGLAVGDL	210.3	Blood glucose >= 180.1	169.2	41.1	404	48HRS

Figure 6: ADBGL with hourly average (PARAMCD=BGLAVGHR) and Daily Average (PARAMCD=BGLAVGDL) with change from baseline based on three different baseline

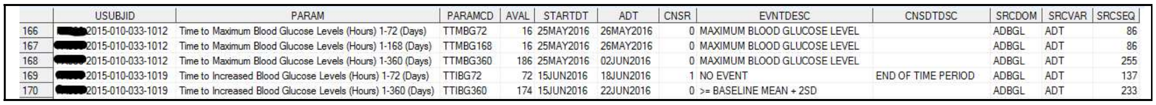
- Furthermore, three other ADaM datasets were created in support of secondary endpoints analyses:
  - Average Blood Glucose Area Under the Effect (ADBGLAUE) [see bullet point 1 described in the “Study Background, Endpoints” section and figure 7].
  - Time-to-events endpoints (ADTTE). Note the data-point traceability vs ADBGL achieved through the use of SRCDOM=ADBGL, SRCVAR=ADT and SRCSEQ being the value in ADBGL.ASEQ [see bullet points 6, 7 and 8 described in the “Study Background, Endpoints” section and figure 8].
  - Categorical endpoints (ADBGCAT). [see bullet point 2 described in the “Study Background, Endpoints” section and figure 9].

. All these datasets were derived from ADBGL being therefore the ADaM Intermediate datasets for the three ADaM datasets.



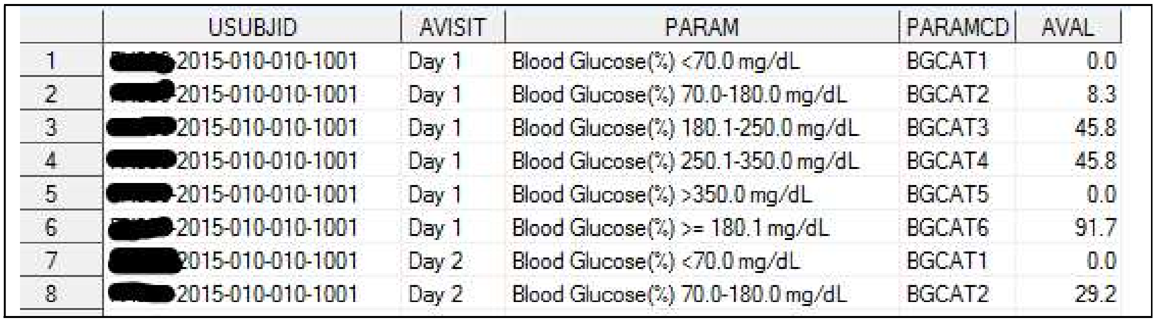
	USUBJID	PARAMCD	PARAM	AVAL
1	[redacted]2015-010-010-1001	BGLAUE1	AUE of Average Blood Glucose Level - Hours -72 to 72	25105.0
2	[redacted]2015-010-010-1001	BGLAUE2	AUE of Average Blood Glucose Level - Hours 0 to 24	5621.5
3	[redacted]2015-010-010-1001	BGLAUE3	AUE of Average Blood Glucose Level - Hours 0 to 48	10669.5
4	[redacted]2015-010-010-1001	BGLAUE4	AUE of Average Blood Glucose Level - Hours 0 to 72	14780.8
5	[redacted]2015-010-010-1001	BGLAUE5	AUE of Average Blood Glucose Level - Hours 0 to 168	29721.7
6	[redacted]2015-010-010-1001	BGLAUE6	AUE of Average Blood Glucose Level - Hours 0 to 360	59573.0
7	[redacted]2015-010-010-1002	BGLAUE1	AUE of Average Blood Glucose Level - Hours -72 to 72	25682.6
8	[redacted]-2015-010-010-1002	BGLAUE2	AUE of Average Blood Glucose Level - Hours 0 to 24	4912.5
9	[redacted]2015-010-010-1002	BGLAUE3	AUE of Average Blood Glucose Level - Hours 0 to 48	9890.0
10	[redacted]2015-010-010-1002	BGLAUE4	AUE of Average Blood Glucose Level - Hours 0 to 72	14794.8

Figure 7: ADBGLAUE with the AUE (PARAMCD=BGLAUEn) by period



	USUBJID	PARAM	PARAMCD	AVAL	STARTDT	ADT	CNSR	EVNTDESC	CNSDTDSC	SRCDOM	SRCVAR	SRCSEQ
166	[redacted]2015-010-033-1012	Time to Maximum Blood Glucose Levels (Hours) 1-72 (Days)	TTMBG72	16	25MAY2016	26MAY2016	0	MAXIMUM BLOOD GLUCOSE LEVEL		ADBGL	ADT	86
167	[redacted]2015-010-033-1012	Time to Maximum Blood Glucose Levels (Hours) 1-168 (Days)	TTMBG168	16	25MAY2016	26MAY2016	0	MAXIMUM BLOOD GLUCOSE LEVEL		ADBGL	ADT	86
168	[redacted]2015-010-033-1012	Time to Maximum Blood Glucose Levels (Hours) 1-360 (Days)	TTMBG360	186	25MAY2016	02JUN2016	0	MAXIMUM BLOOD GLUCOSE LEVEL		ADBGL	ADT	255
169	[redacted]2015-010-033-1019	Time to Increased Blood Glucose Levels (Hours) 1-72 (Days)	TTIBG72	72	15JUN2016	18JUN2016	1	NO EVENT	END OF TIME PERIOD	ADBGL	ADT	137
170	[redacted]2015-010-033-1019	Time to Increased Blood Glucose Levels (Hours) 1-360 (Days)	TTIBG360	174	15JUN2016	22JUN2016	0	>= BASELINE MEAN + 2SD		ADBGL	ADT	233

Figure 8: ADTTE Time to Maximum Blood Glucose Levels (PARAMCD=TTMBGy) and Time to Increased Blood Glucose Levels (PARAMCD=TTIBGy) by period



	USUBJID	AVISIT	PARAM	PARAMCD	AVAL
1	[redacted]-2015-010-010-1001	Day 1	Blood Glucose(%) <70.0 mg/dL	BGCAT1	0.0
2	[redacted]-2015-010-010-1001	Day 1	Blood Glucose(%) 70.0-180.0 mg/dL	BGCAT2	8.3
3	[redacted]2015-010-010-1001	Day 1	Blood Glucose(%) 180.1-250.0 mg/dL	BGCAT3	45.8
4	[redacted]2015-010-010-1001	Day 1	Blood Glucose(%) 250.1-350.0 mg/dL	BGCAT4	45.8
5	[redacted]-2015-010-010-1001	Day 1	Blood Glucose(%) >350.0 mg/dL	BGCAT5	0.0
6	[redacted]-2015-010-010-1001	Day 1	Blood Glucose(%) >= 180.1 mg/dL	BGCAT6	91.7
7	[redacted]2015-010-010-1001	Day 2	Blood Glucose(%) <70.0 mg/dL	BGCAT1	0.0
8	[redacted]2015-010-010-1001	Day 2	Blood Glucose(%) 70.0-180.0 mg/dL	BGCAT2	29.2

Figure 9: ADBGCAT for the “categorical” analysis (PARAMCD=BGCATn)

EXPLAIN WHAT YOU DID THROUGH THE ANALYSIS REVIEWER GUIDE (ADRG)

All the above ADaM datasets and derivation methods were described in the ADaM define.xml. Furthermore, more details were provided in the ADRG to guide the reviewer through the different ADaM datasets so that the reviewer has a clear understanding of the steps followed in the creation of the ADaM datasets in support of the planned analyses.

The following are the standard sections from the PhUSE template that were used to describe what we showed in the previous sections.

3.5 Use of Visit Windowing, Unscheduled Visits, and Record Selection

- Was windowing used in one or more analysis datasets?

Yes in ADBGL for calculating Hourly and Daily means of the Blood Glucose Level (data were originally collected every 5 minutes using CGM device). Derived visits were then used in other ADaM datasets (ADBGCAT, ADBGLAUE and ADTTE).

4.2 Data Dependencies

ADSL was used in the creation of all other analysis datasets.
ADGLRLHR was used in the creation of ADBGL.
ADBGL was used in the creation of ADBGCAT, ADBGLAUE and ADTTE.
ADAE was used in the creation of ADAESUM.

4.3 Intermediate Datasets

ADGLRLHR is an intermediate dataset used to create ADBGL.

4.4 Variable Conventions

ANL01FL was used to flag records copied from SDTM but not used in any analysis e.g. if an observation is not flagged with ‘Y’ then it is not used in the analysis. The flag has been used in the following ADaM data set:

- ADGLRLHR: any observation with a relative hour not falling into scheduled time interval (see Table 2 in the SAP), that is before Hour -72 and after Hour 360.

5.2.3 ADGLRLHR – Intermediate Blood Glucose AD

The ADGLRLHR is an intermediate dataset created from SDTM.ZG where we have derived the relative hours based on blood glucose raw data collected every 5 minutes. It contains one record per subject for every 5 minute blood glucose assessment. Records where ANL01FL=’Y’ are the ones that were used n the analysis (see more details in section 4.4). It is used for creation of ADBGL.

Parameter Code / Name	Description	Usage
BGL Blood Glucose Level (mg/dL)	Contains Blood Glucose level (mg/dL) collected every 5 minutes	This is copied from SDTM.ZG. Relative hours are derived using the assessment date time and treatment date time

5.2.4 ADBGL – Blood Glucose AD

ADBGL contains the primary pharmacodynamics endpoint, that is the Change from Baseline in Daily Average Blood Glucose Level (CHG variable where PARAMCD=BGLAVGDL). In addition ADBGL contains the parameters listed in the table below.
ADBGL contains one record per subject per each derived parameter, per time points. Time points are derived into analysis visits (represented by AVISIT and AVISITN) based on the Relative Hours derived in ADGLRLHR dataset.
Change from baseline (CHG) is used as a dependent variable in some models.
The baseline value (BASE) is used as a covariate in the analysis together with treatment and site (SITEID).

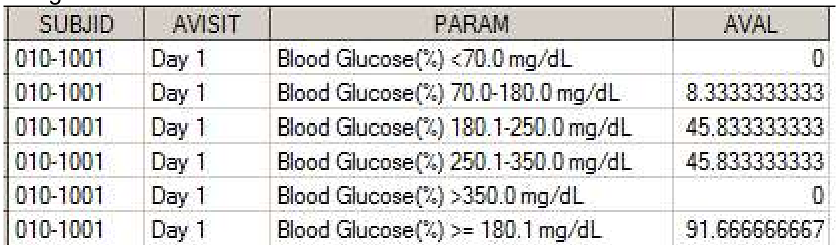
Parameter Code / Name	Description	Usage / Additional Details
BGLAVGHR Hourly Avg. Blood Glucose Level (mg/dL)	Derived parameter that reflects the Hourly Average Blood Glucose Level based on data from ADGLRLHR.	It is derived by taking the average by Relative Hours (ATPTN) from ADGLRLHR where ANL01FL = 'Y'.
BGLAVGDL Daily Avg. Blood Glucose Level (mg/dL)	Derived parameter that reflects the Daily Average Blood Glucose based on the Hourly Average Blood Glucose Level for each day (Day 1-Day 15, Day 1-2, Day 1-3 and so on).	This is the primary pharmacodynamics parameter. It is derived by taking the average by AVISIT where PARAMCD = 'BGLAVGHR'.
BGLAVDLP Daily Avg. Blood Glucose Level (%)	Derived parameter that reflects daily percentage of the total blood glucose for each day (Day 1 - 15)	It was calculated as: (average daily blood glucose / sum of average daily glucose values) * 100
BGLCMDST Cum. Dist. of Blood Glucose (%)	Derived parameter that reflects the cumulative daily percentage of the total blood glucose for each day (Day 1 - 15)	It was derived by adding each daily glucose average (%) to the previous daily averages (%)
BGLMAXD Max. Hourly Avg. Blood Glucose by Day (mg/dL)	Derived parameter that reflects the maximum hourly average blood glucose for each day (Day 1 - 15)	It was derived by taking the maximum value of Hourly blood glucose average for each day
BGLMAXT Max. Hourly Avg. Blood Glucose by Time points (mg/dL)	Derived parameter that reflects the maximum hourly average blood glucose for each time period (Hour 1-24,1-48,1-72,1-168,1-360)	It was derived by taking the maximum value of Hourly blood glucose average for each time period
BGLMXDST Baseline to Maximum Blood Glucose (mg/dL)	Derived parameter that reflects the distance travelled from Baseline to Maximum Blood Glucose for each time period (Hour 1-24,1-48,1-72,1-168,1-360)	It was calculated as: (maximum hourly blood glucose value) - (72-hour baseline blood glucose value)
BGLAVGCV CV - Blood Glucose Level (%)	Derived parameter that reflects the coefficient of variation (CV) of the hourly averages for each time period (Hour 1-24,1-48,1-72,1-168,1-360)	% CV for each subject is derived using the following formula: (SD/mean)*100 for each time period

5.2.5 ADBGLAUE – Blood Glucose Area Under the Effect (AUE) AD
ADBGLAUE contains the endpoints / parameters for the secondary analysis endpoint for the Blood Glucose Area Under the Effect Curve. It is derived from ADBGL by selecting all evaluable post-baseline (AVISITN> 2) hourly average blood glucose derived records (PARAMCD="BGLAVGHR"). Records from ADBGL have been not kept in the dataset.

Parameter Code / Name	Description
BGLAUE1 AUE of Average Blood Glucose Level - Hours -72 to 72 BGLAUE2 AUE of Average Blood Glucose Level - Hours 0-24 ….	Derived using the linear trapezoidal rule ((AVAL(hour)+AVAL(hour+1))/2)*1 The multiplication factor '1' is applied because the parameter BGLAVGHR from ADBGL contains the hourly average for each hour (see SAP Table 2 for more details). All single 'trapezoid' values between the hours specified in the parameter name (i.e. between hour 1 and hour 360 for BGLAVGHR), are then summed.

5.2.6 ADBGCAT – Blood Glucose Categories AD
ADBGCAT was created from ADBGL containing the endpoints/parameters for Blood Glucose Categories for each day (Day 1 - Day 15).
The AVAL was based on the percent of time the hourly average blood glucose was in the following

categories



SUBJID	AVISIT	PARAM	AVAL
010-1001	Day 1	Blood Glucose(%) <70.0 mg/dL	0
010-1001	Day 1	Blood Glucose(%) 70.0-180.0 mg/dL	8.3333333333
010-1001	Day 1	Blood Glucose(%) 180.1-250.0 mg/dL	45.833333333
010-1001	Day 1	Blood Glucose(%) 250.1-350.0 mg/dL	45.833333333
010-1001	Day 1	Blood Glucose(%) >350.0 mg/dL	0
010-1001	Day 1	Blood Glucose(%) >= 180.1 mg/dL	91.666666667

Parameter Code / Name	Description	Usage
BGCAT1 - BGCAT6	1= Blood Glucose(%) <70.0 mg/dL 2= Blood Glucose(%) 70.0-180.0 mg/dL 3= Blood Glucose(%) 180.1-250.0 mg/dL 4= Blood Glucose(%) 250.1-350.0 mg/dL 5= Blood Glucose(%) >350.0 mg/dL 6= Blood Glucose(%) >= 180.1 mg/dL	At each time period, proportions with ≥180.1mg/dL versus <18.01 mg/dL was compared using linear regression (ANCOVA)

5.2.7 ADTTE – Time to Event AD

ADTTE is an analysis dataset following the ADaM TTE Structure (that is based on the BDS structure) supporting Blood Glucose time to event endpoint.

It is derived from ADBGL by selecting Hourly blood glucose average (PARAMCD="BGLAVGHR")

Source data can be traced back to the source analysis dataset using USUBJID and SRCDOM (=ADBGL), SRCVAR (=ADT) and SRCSEQ variables

Parameter Code / Name	Description	Usage
TTIBG72 Time to Increased Blood Glucose Levels (Hours) 1-72 TTIBG360 Time to Increased Blood Glucose Levels (Hours) 1-360	This is defined as the time from administration of study treatment to the first blood glucose level that is ≥ 2 SDs above the 72-hour baseline average blood glucose level (CNSR=0). Subjects without an event were censored (CNSR=1) to the end of the period (Hours 1-72, 1-360)	It is used in the survival model (PROC LIFETEST) with AVAL representing the 'time-to' information and CNSR the censoring flag (1=censor, 0= event).
TTMBG24 Time to Maximum Blood Glucose Levels (Hours) 1-24 TTMBG48 Time to Maximum Blood Glucose Levels (Hours) 1-48 TTMBG72 Time to Maximum Blood Glucose Levels (Hours) 1-72 TTMBG168 Time to Maximum Blood Glucose Levels (Hours) 1-168 TTMBG360 Time to Maximum Blood Glucose Levels (Hours) 1-360	This is defined as the time from administration of study treatment to the maximum hourly average blood glucose level for time period (hours 1-24, 1-48, 1-72, 1-168, 1-360). There was no censoring, as every subject had a maximum value.	It is used in the survival model (PROC LIFETEST) with AVAL representing the 'time-to' information and CNSR the censoring flag (1=censor, 0= event).

CONCLUSIONS

Traceability can be considered to a certain extent an "Art", the art of making complex things simple, clear and transparent. Each of us working with analysis datasets should acquire such an art and apply it in our daily work. Very often programmers and biostatisticians underestimate this aspect, making it difficult to understand for the reviewer or whoever has to take over a study or an analysis task.

In many occasions, while reviewing CDISC packages for some of our clients, we observed severe traceability issues or complete lack of traceability. By looking at some answers/rationale we received we understood some key traceability concepts are either completely ignored or misunderstood. The following is just a simple example from an ADaM CDISC package we did review:

> Observation/Issues Raised: “ADEG (ECG Analysis datasets) with only parameter containing average of triplicates from SDTM”.
>
> Bad Answer: “The original records were not retained in ADEG because they are in SDTM.EG and …..”.
>
> Rationale: although there is no requirement to keep all records from SDTM when creating ADaM datasets, it is a good practice to keep records from SDTM into ADaM when these are the source of records (parameters) derived in ADAM.

By simply following the following four fundamental ADaM principles, our analysis and the steps we followed can be fully understood by anyone looking at what we did:

- Facilitate clear and unambiguous communication and provide a level of traceability
- Be Analysis-Ready
- Be Accompanied by Metadata
- Be useable with commonly available tools

“Splitting” complex processes into smaller pieces such as the concept of intermediate analysis datasets we covered in our paper, can facilitate not only the understanding but also the “reasoning” while approaching such complex processes, making then the solutions easier to apply and eventually to reproduce (validate).

REFERENCES

[1] FDA Study Data Technical Conformance Guide; March 2018

[2] B. Vali, The Importance of CDASH, FDA CDER, CDISC International Interchange, 2013

[3] Analysis Data Reviewer Guide Template; PhUSE

[4] CDISC ADaM Implementation Guidance 1.1, 2016

[5] CDISC Therapeutic Area Data Standards User Guide for Breast Cancer, 2016

[6] CDISC Therapeutic Area Data Standards User Guide for Prostate Cancer, 2017

RECOMMENDED READING

S. Minjoe, T. Petrowitsch, Traceability: Plan Ahead for Future Needs. PhUSE, 2014.

S. Minjoe. Preparing Analysis Data Model (ADaM) Data Sets and Related Files for FDA Submission. PharmaSUG, 2017.

A. Tinazzi, “Lost” in Traceability, from SDTM to ADaM … finally Analysis Results Metadata. CDISC Europe Interchange, 2016.

CONTACT INFORMATION

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

Angelo Tinazzi
Cytel Inc.
Route de Prè-Bois 20
1215 Geneva – Switzerland
angelo.tinazzi@cytel.com
www.cytel.com

Brand and product names are trademarks of their respective companies.