PhUSE EU Connect 2018

Paper DH06

Can You Cut It? Implementing the Data Cut-off

Lewis Meares, Phastar Ltd., London, UK

ABSTRACT

In Oncology studies it is standard practice to cut the data based on a date when a certain number of events have occurred. This is referred to as the data cut-off date and the cut has to be applied with caution. The way to deal with dates after the data cut-off date depends on the type of dates and results that you have. This paper will give guidance for applying the data cut and look into how to deal with different types of dates and results in order to be as safe and correct as possible. We will also go through a macro that can be used to do the data cut while following this guidance.

INTRODUCTION

The data cut-off (DCO) process is used to represent the status of the data at a certain date, known as the DCO date. The DCO process creates a subset of the data so that your datasets contain only data up to and including the DCO date. The date will be chosen by the study team and can be picked in many different ways. The data cut-off is most commonly performed in oncology studies with the date being based on when a certain number of events have occurred (the number of events required would be determined by what will give sufficient power to a study). A data cut-off may occasionally be required for a non-oncology study, for example a yearly summary for a DSUR or a long-term study. In these scenarios the date is most likely to be based on a fixed period of time such as six months or a year.

WHY IS IT NEEDED?

Once we have reached data cut-off date, we need our datasets to contain data collected up to and including that date. Obviously, we need that data to also be clean. This is why data management are usually given a certain number of days after the DCO date to clean the data. However, during the time after the DCO date has occurred and the cleaning process, the database will still be considered open for the investigator to add in any new pre or post DCO data.

Using an example where the DCO date is 1st January 2018, if the data up to and including 1st January 2018 is considered to be fully clean on January 8th, then any assessments or events that occurred between 2nd January 2018 and 8th January 2018 will also be included in the datasets (these records may not be clean either). Therefore, the dates after 1st January need to be removed from the final datasets.

WHEN SHOULD THE DCO PROCESS BE PERFORMED?

The standard dataset creation process is to go from RAW to SDTM and then to ADaM. At some point in this process it is necessary to remove post DCO records to make sure they are not included in any TFLs. The question is when to do it.

Performing the DCO process at the RAW level has the advantage of your SDTM and ADaM datasets having the same version of the data, this will increase traceability in any required submission. The potential disadvantage is that you are manipulating your source data which could lead to confusion about why certain records are not in the RAW data. One solution to this is to create a second set of RAW datasets while keeping the original RAW datasets as they are. You can think of the original RAW datasets as “PRERAW” datasets. This means the process is now:

PRERAW > RAW > SDTM > ADaM

The RAW datasets would be used in your SDTM programs.

Performing the DCO process at the SDTM level is possible, however it is likely to be more complicated than performing it at the RAW level. This is due to raw variables being mapped into separate SDTM datasets. For example, if a patient has a death date after the DCO, you would have to have different approaches for the same

date in different datasets. You would have to make the DTHFL/DTHDTC variables missing in DM, and you're likely to have to delete the record in other SDTMs e.g. CE or DS. If the date was removed at the RAW level this would not be an issue. The death date interdependency between different SDTMs is well-known, but it may be hard to follow all the interdependencies between SDTM datasets, especially if there are different people working on different SDTMs as part of a study team.

Another complication is that as you will be setting records from different CRF pages onto each other, you may want to perform the DCO in different ways for that one SDTM date variable depending on the type of record.

Performing the DCO process at the ADaM level will have the same complications as performing it at the SDTM level. There would also be the issue of traceability between SDTM and ADaM in any required submissions.

After evaluating each approach, performing the DCO process at the RAW level appears to be the best approach. If this approach is chosen, it is recommended to create a PROC COMPARE which will compare the old version and new versions of the RAW datasets.

It is recommended to specify the level the cut-off has been performed in the reviewer's guide.

The rest of this paper will use examples from RAW data to show how to apply the DCO unless stated.

DATE TYPES TO CONSIDER

There are many different types of dates collected in the raw data. Each of those date types needs to be approached in a certain way when implementing the DCO.

In the following examples we will assume a DCO date of 2018-01-01

ASSESSMENT DATES

If an assessment date (mapped to –DTC in SDTM) is after the DCO date, that means the result is after the DCO and therefore the observation should be deleted.

EXAMPLE 1: VITAL SIGNS DATA

Obs. Number	Subject	Visit number	Visit date	Sample Date	Heart Rate
1	1	1	2017-08-28	2017-08-30	80
2	1	2	2018-01-15	2018-01-15	75

The first observation would be kept after the DCO process is performed as both the visit and sample date are before the DCO date. The second observation would be removed in the DCO process as the visit and sample date are after the DCO date.

Companies may decide whether to base the DCO process for assessment dates on either the visit date or the sample date.

START AND END DATES

If a start date (mapped to --STDTC in SDTM) e.g. start date of concomitant medication is after the DCO date, it means the event started after the DCO and that the observation should be deleted. This is regardless of when the event ended.

If an end date (mapped to --ENDTC in SDTM) e.g. end date of concomitant medication is after the DCO date, the observation should be kept but the end date will need to be changed to be missing as the event had not ended at the time of the DCO. The observation should only be deleted if the corresponding start date is after the DCO date.

EXAMPLE 2: START AND END DATES IN CONCOMITANT MEDICATION DATA

Obs. Number	Subject	Concomitant Medication	Start Date of Concomitant Medication	End Date of Concomitant Medication
1	1	INSULIN	2017-11-30	2017-12-31
2	1	PARACETAMOL	2017-11-30	2018-02-15
3	1	VITAMIN D	2018-01-24	2018-02-22

The first observation would have no changes made to it as the start and end dates are both before the DCO date. The second observation would be kept in the dataset as the start date is before the DCO date, but the end date is after the DCO date and therefore the end date would be changed to be missing. The third observation would be deleted from the dataset as the start date is after the DCO date.

BASELINE DATES

Assuming all patients have been recruited before the DCO date, if you have data that is only collected at baseline e.g. medical history or demographics, and the data contains any post DCO dates then you should first query this with data management as this should not happen. If you have patients who are recruited after the DCO date then their information should be deleted.

OTHER DATES

Other Individual dates e.g. date of death; date of withdrawal – If these dates are after the DCO date, the specific observation needs to be deleted.

ADVERSE EVENTS

There are many variables specific to adverse events data that need to be considered when performing the data cut.

SERIOUS ADVERSE EVENTS

If a serious adverse event is recorded with a start date before the DCO date, that observation will definitely be included. However, if you have information that the adverse event started as a non-serious AE and then became a serious AE after the DCO date, you will need to make sure you present the record as a non-serious AE. You will also need to check the reason for serious AE variables to make sure they are all missing.

EXAMPLE 3: SAE DATA

Adverse event	AE Start Date	Date AE Became Serious	Serious AE	Results in Death	Requires or Prolongs Hospitalisation	Is Life Threatening	Other Medically Important Serious Event
Nausea	2017-12-15	2018-01-17	Y		Y		

In example three above you will need to make the Serious AE and Requires or Prolongs Hospitalisation flags missing as the AE became serious after the DCO date.

OUTCOME OF ADVERSE EVENT

Once an adverse event has ended then the outcome is that the adverse event is RECOVERED/RESOLVED or RECOVERED/RESOLVED WITH SEQUELAE. If the end date is after the DCO date, then as the end date will be made to be missing it means the AE can no longer considered to be resolved. It is recommended to take a conservative approach here and say the outcome is NOT RECOVERED/NOT RESOLVED, as opposed to RECOVERING/RESOLVING.

If the outcome of the adverse event is FATAL, you need to check if the patient's date of death was before or after the DCO date. If the date of death was after the DCO, the outcome needs to be changed to be NOT RECOVERED/NOT RESOLVED. It will be easier if this part of the DCO process is done before you perform the

DCO process for the death date. Otherwise it could be confused with a data issue if the death date is missing due to it being deleted for being post the DCO date.

ACTION TAKEN

Whichever action is taken you need to look at the dose data you have to check if that action taken occurred after the DCO date. If the action was taken after the DCO date, you need to change the action taken to be the most severe action taken in the patients dosing data before the DCO date e.g. DRUG INTERRUPTED. Before applying the change, you need to be sure that the action taken in the dosing data was definitely due to that adverse event. If no actions were taken for that adverse event before the DCO date, the action taken would be DOSE NOT CHANGED.

In example four below the patient has had their dose reduced due to an adverse event, however you can see from the dosing data that the dose wasn't reduced until after the DCO. Therefore, the action taken needs to be changed to an earlier action taken. The next most severe action taken before the DCO date is a drug interruption on 2017-12-16. But because the reason for the drug interruption was that the subject forgot to take the dose, that is unrelated to the adverse event so the action taken should be changed to be DOSE NOT CHANGED.

EXAMPLE 4.1: AE DATA

Adverse event	AE Start Date	AE End Date	Action Taken
Headache	2017-12-15	2018-02-30	DOSE REDUCED

EXAMPLE 4.2: DOSING DATA

Dose	Frequency	Dose Start Date	Dose End Date	Action Taken	Reason for Action Taken
80	BID	2017-11-30	2017-12-15	DOSE NOT CHANGED	
0		2017-12-16	2017-12-16	DRUG INTERRUPTED	Subject forgot to take dose
80	BID	2017-12-17	2018-01-16	DOSE NOT CHANGED	
40	BID	2018-01-17	2018-02-30	DOSE REDUCED	Adverse Event

If the action taken is DRUG WITHDRAWN, it's likely that there won't be an observation in the dosing data confirming when that took place. In this situation you should look at whether the patients last day of dose was before or after the DCO date. If the last dose was after the DCO date, that shows the drug was not withdrawn until after the DCO date and you need to follow the guidance above on changing their action taken.

EXAMPLE 5.1: AE DATA

Adverse event	AE Start Date	AE End Date	Action Taken
Nausea	2017-12-15	2018-02-30	DRUG WITHDRAWN

EXAMPLE 5.2: DOSING DATA

Dose	Frequency	Dose Start Date	Dose End Date	Action Taken	Reason for Action Taken
80	BID	2017-11-30	2017-12-14	DOSE NOT CHANGED	
0		2017-12-15	2018-12-19	DRUG INTERRUPTED	Adverse Event
80	BID	2017-12-20	2018-01-03	DOSE NOT CHANGED	Adverse Event

In example five above the action taken is DRUG WITHDRAWN, however you can see from the dosing data example that the patient had taken the drug beyond the DCO date therefore the drug was not withdrawn before the DCO. This means we need to change the action taken from DRUG WITHDRAWN to the next most severe action taken before the DCO which in the example above is DRUG INTERRUPTED.

END FLAGS

As discussed earlier in the paper, if an end date is after the DCO date then that end date should be made to be missing. Another thing to consider is if there is an end flag associated with the end date. If for example a concomitant medication has an associated flag saying the medication has ended, but the end date of the concomitant medication is after the DCO, the end flag needs to be changed to indicate the medication is ongoing.

EXAMPLE 6.1: CONCOMITANT MEDICATION DATA – PRE DCO PROCESS

Has the medication ended?	End date of concomitant medication
Yes	2018-01-13

EXAMPLE 6.2: CONCOMITANT MEDICATION DATA – POST DCO PROCESS

Has the medication ended?	End date of concomitant medication
No	-

In SDTM this would be shown by either --ENRF or – ENRTPT being equal to “ONGOING”. The same variables can be carried through to ADaM as well.

PARTIAL DATES

Occasionally you may only have the year and month parts of a date, and if the month is in the same month as the DCO date then you can’t be sure whether it was actually before or after the data cut-off date. The same issue applies if you only have the year and if the year was the same year as the DCO date.

To solve this issue, the recommendation is to first apply the PhUSE guidance on imputing partial start and end dates. After imputing the start and end dates you can then apply the guidance in this paper.

If performing the DCO at RAW or SDTM level DO NOT add in the imputed day or month. At the ADaM level the imputation can be added into the ASTDT and AENDT variables with ASTDTF and AENDTF appropriately populated to indicate the level of imputation.

PARTIAL START DATES

Below are examples of partial adverse event start dates and if they would be included after the DCO process is applied. For these examples, we will assume the date of the patient’s first dose is in 2017 and the PhUSE guidance is being followed.

EXAMPLE 7: PARTIAL AE START DATES

AE start date	AE start date after imputation rules applied	Included after DCO process applied?
2018-01	2018-01-01	Yes
2018	2018-01-01	Yes

PARTIAL END DATES

Below are examples of partial adverse event end dates and if the partial end date will be populated in the RAW or SDTM dataset after the DCO process is applied.

EXAMPLE 8: PARTIAL AE END DATES

AE end date	AE end date after imputation rules applied	Date populated after DCO process applied?
2018-01	2018-01-31	No
2018	2018-12-31	No

OTHER PARTIAL DATES

If you have a partial date that isn't a start or end date, for example a date of collection, and the year or month is the same as the DCO then there isn't a clear imputation rule. It is recommended to include the date as it is a more conservative approach to say the event, such as a death, occurred before the DCO.

USING THE DCO DATE BEYOND THE CUT-OFF PROCESS

Once you've performed the cut-off, your likely to still need to use the DCO date in your dataset programming to resolve missing dates that will cause issues further down the line in your TFL programming.

For example, in ADAE, when calculating the variable TRTEMFL to indicate whether an adverse event is on treatment, you ideally want an end date of treatment to make the derivation easier. In the example below, we have no treatment end date as the patient has not ended study treatment yet.

EXAMPLE 9.1: ADAE

AEDECOD	ASTDT	TRTSDT	TRTEDT
Nausea	28MAY2017	30APR2017	

We can tell that this adverse event is on treatment, however with a missing end date it is a little harder to program as if your code follows the standard derivation for TRTEMFL found in ADaM of:

```
If ADSL.TRTSDT≤ASTDT≤ADSL.TRTEDT + x days then TRTEMFL='Y'
```

Using this derivation, then this example will have TRTEMFL as missing as it would not be considered to be ≤TRTEDT due to TRTEDT being missing.

The way to resolve this would be to have the DCO date as your treatment end date so that the TRTEMFL derivation will resolve to Y.

EXAMPLE 9.2: ADAE WITH DCO DATE ADDED

AEDECOD	ASTDT	TRTSDT	TRTEDT	TRTEMFL
Nausea	28MAY2017	30APR2017	**01JAN2018**	Y

You may want to populate DM.RFENDTC with the DCO date as well for the same situation as above, that would also help to remove any errors from Pinnacle 21 such as the ones below:

- RFENDTC is not provided for a randomized subject
- Value for --ENRF is populated, when RFENDTC is NULL

Another example would be in the Subject Elements (SE) dataset where you would ideally want SEENDTC to be populated each time to be able to allocate the EPOCH variable easily to the other SDTMs.

If you use the DCO in the examples above it means you will to have to use the DCO date in multiple programs. To avoid not updating the date in one of the programs by mistake, you may want to consider having a macro variable in a separate SAS macro. The macro would then be called in each program it is needed in. The macro variable would need to be defined as a global macro. See an example below:

```
%macro cutoff;

  %global dcodate;

  %let dcodate = %str(2018-01-01);

%mend cutoff;
```

DATES INCLUDING TIMES

If the variable you are applying the cut-off to only contains the year, month and day then you can simply say where the variable is less than or equal to the DCO date. However, if the variable includes the time as well then this wouldn't work. This is because SAS® defines having no date as being the same as T00:00:00

Consider the following example of two variable dates where the data cut-off date is 2018-01-01 with no time included:

Variable date	Variable date according to SAS	Date included after DCO applied?
2018-01-01	2018-01-01T00:00:00	Yes
2018-01-01T12:00:00	2018-01-01T12:00:00	No

To make sure the 2018-01-01T12:00:00 date is included you need to make the DCO date 2018-01-01T23:59:59

If the date variable is numeric you will have to find the numeric value that the data cut-off date corresponds to. A simple way to do this would be to write the following code:

```
%let DCO=2018-01-01T23:59:59;

Data ___;

  Set ___ (where=(date<=”&DCO.”dt));

Run;
```

Of course, if you wanted to remove records after a certain time you could change the time part of the macro variable. e.g. for 2pm you could write 2018-01-01T14:00:00.

EXAMPLE CODE

The following code is based on applying the DCO process at the RAW data level, but it can be easily adapted to SDTM level.

EXAMPLE 10: DELETING RECORDS

The code below has an array which will contain all the dates that you want to use to be able to delete records from a particular dataset. The reason that the code contains an array is so that you can check multiple dates in the dataset at the same time. If you didn't have an array you would have to create separate data steps for each date in the dataset. The code would become more complicated as you would need to make sure you are creating a new version of the dataset each time to make sure you aren't overwriting the dataset created in the previous data step.

```
%macro del

(ds= /*input dataset*/

,dat= /*date variables from ds dataset to compare to DCO date*/);

  data raw.&ds.;

    set preraw.&ds.;

    array orig_dat &dat.;

      do over orig_dat;

        if substr(orig_dat,1,10)>"&DCO." then delete;

      end;

  run;

  proc compare base=preraw.&ds. compare=raw.&ds.;
  run;

%mend del;

%del(ds=labraw, dat=labdat);
```

EXAMPLE 11: START AND END DATES

The code below has two separate sections. The first section is for any start dates in your dataset. Similar to the code above if you classify any dates as start dates it will delete any records where the start date is after the DCO date.

The second section of the code is for any end dates in your dataset. This time if any dates that you have classified as an end date are after the DCO date it sets the end date to be blank.

```
%macro startend

(ds= /*input dataset*/

,sdat=none /*start dates to compare against DCO date*/

,edat=none /* end dates to compare against DCO date*/);

  data raw.&ds.;

    set preraw.&ds.;

      %if &sdat. ne none %then %do;

        array a_sd &sdat.;

        do over a_sd;

          if substr(a_sd,1,10)>"&DCO." then delete;

        end;

      %end;

      %if &edat. ne none %then %do;

        array a_ed &edat.;

        do over a_ed;

       if substr(a_ed,1,10)>"&DCO." then a_ed="";

        end;

      %end;

    run;

    proc compare base=preraw.&ds. compare=raw.&ds.;
    run;

%mend startend;

%startend(ds=ae,sdat=aest,edat=aeen);
```

CONCLUSION

In summary, there are many different things to consider when performing the data cut-off. The first thing to consider is at what stage to do the DCO process. This paper recommends to perform the DCO process at the raw stage and to create a new set of raw datasets, with the original raw datasets to be considered as “preraw” datasets. The main dataset specific issues are in the adverse events dataset and there are multiple variables within the AE dataset to check such as the action taken and outcome variables. If the date variables contain times you’ll need to add “T23:59:59” to the DCO date. The DCO date is likely to be needed after the DCO process as well, because of that this paper recommends storing the DCO date in a separate macro which can be called in any program.

REFERENCES

[1] https://www.phusewiki.org/wiki/index.php?title=Imputing_Partial_Dates

ACKNOWLEDGMENTS

I would like to thank Barbara Ramambason, Kevin Kane and Michael Fitzgerald from Phastar Ltd. for their review and advice on this paper. I would also like to thank AstraZeneca for their support in letting me do this paper.

CONTACT INFORMATION

Your questions are welcomed and encouraged. Please contact the author at:
Lewis Meares
Phastar Ltd.
Unit 2A, 2 Bollo Lane,
London
W4 5LE
Email: lewis.meares@phastar.com