PharmaSUG China 2021 - Paper DM-023

Implementation of Data Cut-Off in Analysis of Oncology Clinical Trials

Shang Shi; Weibin Cai; Zhiping Yan, Dizal Pharma, Beijing, China

ABSTRACT

In Oncology studies, it is common practice to cut the data based on a date when a certain number of events have occurred or when a pre-specified milestone is reached. Data Cut-off (DCO) plays a crucial part in support interim analysis, since it has a major impact on the interpretation of trial results. The paper describes the methodology of performing DCO on collected RAW data and gives detailed instructions of each step of the DCO process. The process starts with a SAS macro to automatically generate an Excel file of DCO Specification and a Word file containing DCO rules for each RAW data set, based on RAVE Architect Loader Specification (ALS) file. And then, a data cut macro is developed to read the Excel Specification and apply the DCO for RAW data to get the post-DCO data sets. Finally, a self-check program has been developed to exam whether the post-DCO data sets indeed follow the Specification. The cut-off process has been integrated into Dizal-iSCP (Dizal-Integrated Statistical Computing Platform), which streamlines the DCO process.

INTRODUCTION

In this paper, DCO is considered as the process of restricting data up to a specific date for analysis, and the pre-defined date is referred as DCO date. The DCO plays a crucial part in support interim analysis for oncology clinical trials. In order to perform formal interim analysis, the Data Management (DM) team would usually be offered several weeks after cut-off date to perform data cleaning. During which, the Electronic Data Capture (EDC) system is still open to the investigators to add records which may start or occur after the cut-off date. By the time data cleaning is completed, data will be extracted from EDC, and the RAW data might contain subjects and records that should not be included for the interim analysis.

The DCO creates a subset of data which contains only data collected on or before the DCO date (data kept), and another subset contains data after the DCO date (data removed). The DCO date is decided by study team for analysis, regulatory or other purposes.

The implementation of DCO is most commonly performed in oncology studies when a) a certain number of events occurred or b) a study milestone or specified duration of follow-up is reached[1], including but not limited to following situations:

1. Supporting sample size-recalculation for adaptive trial design
2. Preparing supportive Table, Listing, Figures (TLFs) for regulatory activity: Development Safety Update Report (DSUR), Investigational New Drug (IND) application, Breakthrough Therapy Designation (BTD) application using ongoing trial data, New Drug Application (NDA) and so on.

The application of DCO must be carried out carefully, for it has a major impact on the interpretation of trial results. For oncology study, especially in early phase when there are limited number of patients enrolled, incorrectly implementation of DCO could lead to difference in endpoint assessment.

WHEN SHOULD THE DCO BE PERFORMED AND WHY

The standard data set creation process is to go from RAW data to SDTM, and to ADaM. The application of DCO could be carried out at different stages of data process, either on

a) RAW data: data collected from the CRF and extracted from EDC, or provided by external sources, such as central lab or
b) SDTM data sets: CDISC standards for study data tabulation.

Both of the methods have advantages and disadvantages. Performing the DCO on RAW data means that the SDTM and ADaM data set would be generated from the same post-DCO data, which would greatly increase traceability[1][1]. However, manipulating source data could lead to confusion of missing original records[3].

On the other hand, performing the DCO on SDTM data sets would avoid coding on source data, but it has been proven to be more difficult. Because, for majority SDTM domains, records are set from multiple RAW data sets with multiple timing variables, and some of domains are related and from a CRF page. For example, firstly, date variable is mapped into one domain while not needed in another, e.g. Date of Death is required in DD domain, meanwhile, the date of Death and Death Flag is included in DM domain. Secondly, the change of SDTM data would cause the change of the corresponding supplemental and relationship domain.

Taking all of the above into consideration, we believe that performing the DCO on RAW data would be more appropriate. To avoid confusion on missing source data, the DCO process state in this article split the RAW data into two sets of data with records kept for analysis and records removed, the pre-DCO data set and post-DCO data set are stored in different library.

In the example given below, pre-DCO data sets are stored in library LPTEDC, and post-DCO data sets are stored in library LPTEDCC with an additional letter C append to the end. Data set with records kept for analysis will have the same name as that of its input, while data set hold the records removed will have the same data set name followed by '_CUT' at the end. Through this method, the RAW data will be preserved, it will be cut into two data sets, and the set of post-DCO data should be identical with pre-DCO data.

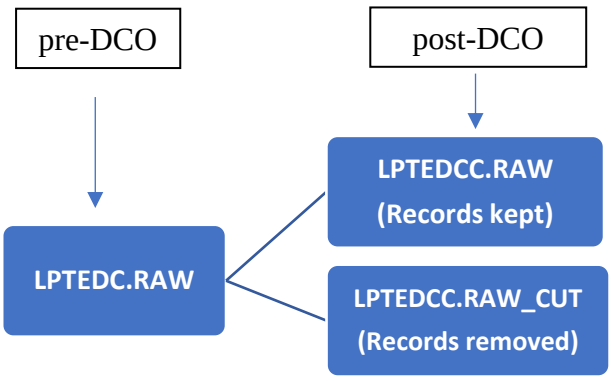

Figure 1 Example of DCO data set storage

HOW WOULD THE DATA BE CUT?

Level 1: Subject Level

Subject with informed consent obtained later than the cut-off date will be excluded from all collected data. Subject level cut will be applied on all data sets first, then move on to record level cut.

Level 2. Record Level

After applying subject level cut, records with assessment date, event date or intervention date after the cut-off date will be removed, following the cut-off rules listed below.

As stated in previous section, for an ongoing study, it is inevitable that some data will be unclean. It is also common for the date variable to have missing or unknown component. When encountered uncertainty about partial date, the general principle is to include as many records as possible[2].

- For Partial date, only compare the available date part with the same part of cut-off date

  (1) If year and month are available: remove record if the year-month is later than the year-month of the cut-off date. For example, in a situation that the cut-off date is April 15, 2021 and the start date of certain Concomitant Procedure (CM) is MAY2021 with missing day information, the date is compared with only the year and month part of the cut-off date, which is APR2021, is later than the year-month of DCO. Therefore, the record will be removed for further analysis.

  (2) If only year is available, remove record if the year is later than the year of the cut-off date. Continue with the example above, if the start date of certain Adverse Event (AE) is 2021, the date is compared with only the year part of the cut-off date, which is also 2021. Therefore, the record will not be removed.

- For completely missing date, follow general principle, keep the record.

For the complex cutting logic, such as overall response of tumor assessment, there is no need to perform record level cut, because the date is unavailable, or the date of scan does not determine the response. For example, in Dizal Data Management practice, tumor assessment might be collected in multiple data sets and one visit could cross several dates. If the overall response is Progression Disease, data of overall assessment is usually mapped to the earliest date, while if subject is a responder, the overall assessment date would be mapped to the latest. In this case, the cut-off process should be completed in the program of SDTM development. Another case is AE start before the DCO and toxicity grade change after the DCO, our practice is to keep all toxicity grade change information in the AE domain regardless of whether the date of toxicity change is before the DCO or not as the AE action or outcome should be revised if the grade change after DCO is eliminated. This will be explained in cSDRG and the complex logic will be handled during the ADaM development.

DATA CUT-OFF PROCESS

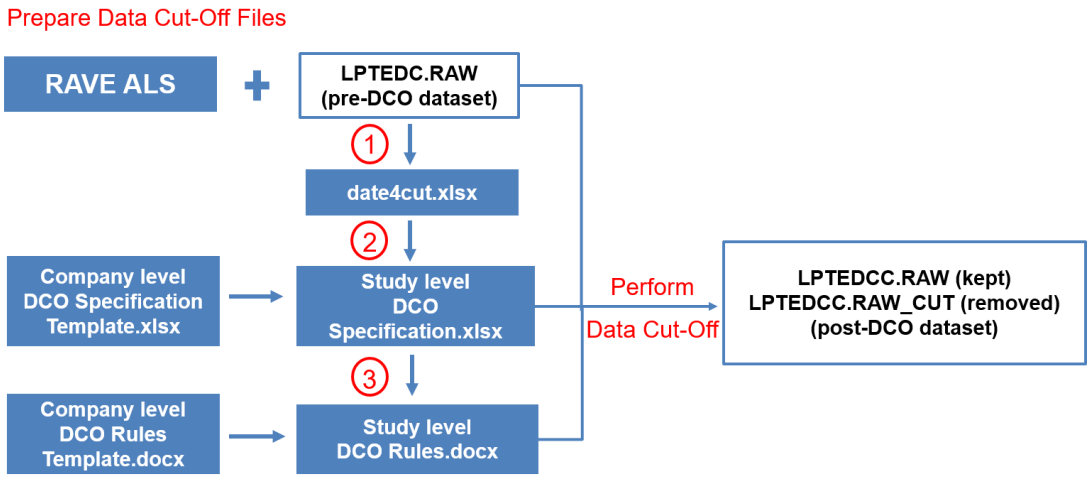

Figure 2 Flow chart of Data Cut-Off Process

Prepare Data Cut-Off Files

Step 1. Generate date4cut.xlsx

1) Refer to the information provided by RAVE ALS, use ***%u_edc_date4cut*** to collect all date variables and related information from RAW data, and export to an Excel file date4cut-temp. The structure of the file is one row per data set per date variable.

A	B	C	D	E	F
Lib	DsName	DsLabel	VarName	VarLabel	Removed
LPTEDC	MH	Medical History	MHSTDAT_RAW	Start date (DD/MMM/YYYY) (RAW)	
LPTEDC	MH	Medical History	MHENDAT_RAW	End date (DD/MMM/YYYY) (RAW)	
LPTEDC	NC	Next Cycle Status			
LPTEDC	OVELRESP	Overall Response (Lugano Assessment)	TR_DT_RAW	Date of assessment (DD/MMM/YYYY) (RAW)	
LPTEDC	OVERDOSE	Overdose	OD_SDAT_RAW	Start date of overdose (DD/MMM/YYYY) (RAW)	
LPTEDC	OVERDOSE	Overdose	OD_EDAT_RAW	Stop date of overdose (DD/MMM/YYYY) (RAW)	
LPTEDC	PATHGEN	Pathology	FDMD_DAT_RAW	Date of initial pathological diagnosis (RAW)	
LPTEDC	PE	Physical Examination	PEDAT_RAW	Date of assessment (DD/MMM/YYYY) (RAW)	
LPTEDC	PREG	Pregnancy Test	SAMP_DAT_RAW	Date of sample collection (DD/MMM/YYYY) (RAW)	
LPTEDC	PSTAT	ECOG Performance Status	PSTATDAT_RAW	Date of assessment (DD/MMM/YYYY) (RAW)	
LPTEDC	PULM	Pulmonary Function Test	ASM_DAT_RAW	Date of assessment (DD/MMM/YYYY) (RAW)	
LPTEDC	RADIRESP	CT/MRI Based Response_LUGANO			
LPTEDC	SPCBEDB	Blood Sample for Mutation Detection	SPECDAT_RAW	Date of Collection (DD/MMM/YYYY) (RAW)	
LPTEDC	SPCBEDT	Tumor Tissue Collection	BMDAT_RAW	Date of Collection (DD/MMM/YYYY) (RAW)	
LPTEDC	SPCBEDT	Tumor Tissue Collection	BMRLTDAT_RAW	Date of Biopsy/Resection (RAW)	

Figure 3 Example of date4cut-temp.xlsx

Column A to E contains information collected from ALS. If a raw data set does not contain specific date or date variable is not applicable, then column D&E will be blank.

Column F is named as REMOVE indicating whether the date variable shall be removed for the next. If the date variable is not needed for DCO, fill the column with “Y”, else leave it blank. The decision is made by the statistical programmer and consult with medical and statistician if encounter any uncertainty.

2) Based on the temporary file, use ***%u_edc_date4cut*** to generate date4cut-final.xlsx.

A	B	C	D	E	F
LIB	DSNAME	DSLABEL	DTVAR_1	DTVAR_2	DTVAR_3
LPTEDC	MH	Medical History	MHSTDAT_RAW	MHENDAT_RAW	
LPTEDC	NC	Next Cycle Status			
LPTEDC	OVELRESP	Overall Response (Lugano Assessment)	TR_DT_RAW		
LPTEDC	OVERDOSE	Overdose	OD_SDAT_RAW	OD_EDAT_RAW	
LPTEDC	PATHGEN	Pathology	FDMD_DAT_RAW		
LPTEDC	PE	Physical Examination	PEDAT_RAW		
LPTEDC	PREG	Pregnancy Test	SAMP_DAT_RAW		
LPTEDC	PSTAT	ECOG Performance Status	PSTATDAT_RAW		
LPTEDC	PULM	Pulmonary Function Test	ASM_DAT_RAW		
LPTEDC	RADIRESP	CT/MRI Based Response_LUGANO			
LPTEDC	SPCBEDB	Blood Sample for Mutation Detection	SPECDAT_RAW		
LPTEDC	SPCBEDT	Tumor Tissue Collection	BMDAT_RAW	BMRLTDAT_RAW	

Figure 4 Example of date4cut-final.xlsx

The date4cut-final.xlsx only includes date variable needed for the DCO. The structure is one row per RAW data set, which is easier to differentiate data sets with different number of date variables. For the data set that does not contain any date variable, such as Next Cycle Status (NC) and CT/MRI Based Response_LUGANO (RADIRESP), column DTVAR_1 will be left blank. In Figure 4, there are maximum 2 date variables included for each row, however, there might be more in real cases.

Step 2. Generate Study-Level-DCO-Specification.xlsx

Use date4cut-final.xlsx, Company level DCO Specification Template and call ***%u_edc_cutspecs*** to generate Study-Level-DCO-Specification.xlsx.

A	B	C	D	E	F
LIB	DSNAME	DSLABEL	DTVAR	SUBJECTID	Comment
LPTEDC	AE	Adverse Events	AESTDAT_RAW		
LPTEDC	ASMPERF	Brain MRI Scan	ASM_DAT_RAW		
LPTEDC	BIOCLINICA	BIOCLINICA	EXAMDAT		
LPTEDC	BMAB1	Bone Marrow Aspiration and Biopsy	ASMDT_RAW | ASPRDAT_RAW		
LPTEDC	BMAB2	Bone Marrow Aspiration and Biopsy_Baseline	ASMDT_RAW | ASPRDAT_RAW		
LPTEDC	CAPRX1	Prior Cancer Therapy	CXSDAT_RAW		
LPTEDC	CAPRX2	Cancer Therapy after end of study treatment	CXSDAT_RAW		
LPTEDC	CAPRXR1	Prior Radiotherapy	CXRDSDAT_RAW		
LPTEDC	CAPRXR2	Post IP Discontinuation Radiotherapy	CXRDSDAT_RAW		
LPTEDC	CAPRXR3	Concomitant Radiotherapy	CXRDSDAT_RAW		
LPTEDC	CM	Prior and Concomitant Medications	CMSTDAT_RAW		
LPTEDC	CO	Comment	CODTC	SUBJID	
LPTEDC	CONPRO	Concomitant Procedures	PROSDAT_RAW		
LPTEDC	CONSENT	Informed Consent	MCONSDAT_RAW		
LPTEDC	CONSWD	Withdrawal of Informed Consent	MCWDDAT_RAW		
LPTEDC	CTMSINTG	CTMS Integration (Hidden)	NONE		no need to cut since of system data
LPTEDC	DEATH	Death	DTH_DAT_RAW		
LPTEDC	DECG	Central ECG	EGDAT_RAW		
LPTEDC	DM	Demographics	NONE		cut at subject level only
LPTEDC	DOSDISC	Discontinuation of Investigational Product	IPDCDDAT_RAW		
LPTEDC	DS	Disposition	DSSTDAT_RAW		
LPTEDC	EF	ECHO/MUGA	EFDAT_RAW		
LPTEDC	EG	ECG	EGDTC	SUBJID	

Figure 5 Example of Company-Level-DCO-Specification-Template.xlsx

Company Level DCO Specification Template contains all available modules from Dizal data collection standards document, and it is regularly updated responding to the needs of different studies.
Column A, B and C included SAS library the RAW data stored in, data set name and data set label, respectively. They are fixed for all studies.

Column D (DTVAR) indicates the date variable and the logical expression used for the DCO. Ampersand ('&') indicates logical conjunction expression (AND), and vertical bar ('|') indicates logical disjunction expression (OR). The operation sequence of logical expression is consistent with the arithmetic expression, 'AND' has higher priority than 'OR'.

For example:

	LIB	DSNAME	DSLABEL	DTVAR
Example 1	LPTEDC	VS	Vital Signs	(DATE1 & DATE2) | (DATE1 & DATE3)
Example 2	LPTEDC	VS_1	Vital Signs	DATE1 | DATE2 & DATE3
Example 3	LPTEDC	VS_2	Vital Signs	DATE1 | DATE2 | DATE3
Example 4	LPTEDC	VS_3	Vital Signs	DATE1 & DATE2 & DATE3

Example 1: data will be removed if (DATE1 > cut-off and DATE2 > cut-off) or (DATE1 > cut-off and DATE3 > cut-off)
Example 2: data will be removed if (DATE1 > cut-off) or (DATE2 > cut-off and DATE3 > cut-off)
Example 3: data will be removed if DATE1 > cut-off or DATE2 > cut-off or DATE3 > cut-off
Example 4: data will be removed if (DATE1 > cut-off and DATE2 > cut-off and DATE3 > cut-off)

Column E (SUBJECTID) is used as indicator for ID variable. Data collected from EDC normally share common identification (ID) variable - SUBJECT. However, ID variable from external data may be named differently, such as SUBJECTID, SUBJID, PATIENT_ID, PARTICIPANT_ID, etc. The ID variable is used during the cut-off macro, it's important to point out the correct variable when it is not default. If RAW data does not contain SUBJECT, the macro will search all variables for potential ID variables and populated it for Column E in the Study Level Spec. If there is more than one available ID, all of them will be listed for further validation.

Column F (COMMENT) is used to provide additional note for reviewer, it will be placed in the cut-off rule document.

A	B	C	D	E	F
LIB	DSNAME	DSLABEL	DTVAR	SUBJECTID	COMMENT
LPTEDC	AE	Adverse Events	AESTDAT_RAW		
LPTEDC	ASMPERF	Brain MRI Scan	ASM_DAT_RAW		
LPTEDC	BIOCLINICA	BIOCLINICA	EXAMDAT		
LPTEDC	BIOPSY	Biopsy	BIODAT_RAW		
LPTEDC	BMAB1	Bone Marrow Aspiration and Biopsy	ASMDT_RAW | ASPRDAT_RAW		
LPTEDC	BMAB2	Bone Marrow Aspiration and Biopsy_Baseline	ASMDT_RAW | ASPRDAT_RAW		
LPTEDC	CAPRX1	Prior Cancer Therapy	CXSDAT_RAW		
LPTEDC	CAPRX2	Cancer Therapy after end of study treatment	CXSDAT_RAW		
LPTEDC	CAPRXR1	Prior Radiotherapy	CXRDSDAT_RAW		
LPTEDC	CAPRXR2	Post IP Discontinuation Radiotherapy	CXRDSDAT_RAW		
LPTEDC	CAPRXR3	Concomitant Radiotherapy	CXRDSDAT_RAW		
LPTEDC	CM	Prior and Concomitant Medications	CMSTDAT_RAW		
LPTEDC	CO	Comment	CODTC	SUBJID	
LPTEDC	CONPRO	Concomitant Procedures	PROSDAT_RAW		
LPTEDC	CONSENT	Informed Consent	MCONSDAT_RAW		
LPTEDC	CONSENT2	Informed Consent for pSTATs	MCONSDAT_RAW		
LPTEDC	CONSWD	Withdrawal of Informed Consent	MCWDDAT_RAW		
LPTEDC	CTMSINTG	CTMS Integration (Hidden)	NONE		no need to cut since of system data
LPTEDC	DEATH	Death	DTH_DAT_RAW		
LPTEDC	DECG	Central ECG	EGDAT_RAW		
LPTEDC	DM	Demographics	NONE		cut at subject level only
LPTEDC	DOSDISC	Discontinuation of Investigational Product	IPDCDDAT_RAW		
LPTEDC	DS	Disposition	DSSTDAT_RAW		
LPTEDC	EF	ECHO/MUGA	EFDAT_RAW		
LPTEDC	EG	ECG	EGDTC	SUBJID	
LPTEDC	ENROL	Enrollment	NONE		cut at subject level only

Figure 6 Example of Study-Level-DCO-Specification.xlsx

The ***%u_edc_cutspecs*** macro check the number and name of date variable in the date4cut-final.xlsx and the number and name of date variable in the DCO-specification-template.xlsx. If they are matched, then carry the content from template. If not, write variable name from date4cut-final.xlsx and highlight the cell, indicating requirement of further validation or update.

Step 3. Generate Study Level DCO rules.docx

After finish validating and updating all highlighted cells, use ***%u_edc_cutrule*** to combine Study-Level-DCO-Specification.xlsx and Company level DCO rules template.docx and generate a word file of Study Level DCO rules. The body of document includes the general information regarding DCO and its methodology. The details in Appendix 1 are matched with variables listed in DCO Specification.

Appendix 1: Cut-off Details for each raw dataset

Dataset	Apply Cut?	Description	Cut-off rules
ASMPERF	Yes	Brain MRI Scan	Remove record if ASM_DAT_RAW > cut-off date
BIOCLINICA	Yes		Remove record if EXAMDAT > cut-off date
BIOPSY	Yes		Remove record if BIODAT_RAW > cut-off date
BMAB1	Yes	Bone Marrow Aspiration and Biopsy	Remove record if ASMDT_RAW | ASPRDAT_RAW > cut-off date
BMAB2	Yes	Bone Marrow Aspiration and Biopsy_Baseline	Remove record if ASMDT_RAW | ASPRDAT_RAW > cut-off date
CAPRX1	Yes	Prior Cancer Therapy	Remove record if CXSDAT_RAW > cut-off date
CAPRX2	Yes	Cancer Therapy after end of study treatment	Remove record if CXSDAT_RAW > cut-off date
CAPRXR1	Yes	Prior Radiotherapy	Remove record if CXRDSDAT_RAW > cut-off date
CAPRXR2	Yes	Post IP Discontinuation Radiotherapy	Remove record if CXRDSDAT_RAW > cut-off date
CAPRXR3	Yes	Concomitant Radiotherapy	Remove record if CXRDSDAT_RAW > cut-off date
CM	Yes	Prior and Concomitant Medications	Remove record if CMSTDAT_RAW > cut-off date
CO	Yes		Remove record if CODTC > cut-off date

Figure 7 Example of Study Level DCO Rules.docx

Edit on the draft document to add study specific information and addition note in appendix 1, e.g., tumor assessment related data sets for oncology trials. The file will help the regulatory authorities and other functions within the study team to better understand the DCO rules.

Perform Data Cut-Off

After preparing all essential document, performing DCO will be relatively straightforward.

1. Use macro ***%m_sdtm_dscut*** to cut the RAW data based on cut-off rules stated in Study-Level-DCO-Specification.xlsx.
2. Use macro ***%u_edc_cutchk*** to Check the post-DCO data sets based on following rules:
   1) Subject enrolled before DCO date should be included;
   2) DCO date of the records kept are up to or including cut-off date;
   3) DCO date of the records removed are after cut-off date.

Integrated into Dizal-iSCP

To further streamlines the DCO process, it was integrated into Dizal-iSCP (Dizal-Integrated Statistical Computing Platform). Dizal-iSCP is an interactive programming interface developed by Dizal infrastructure team. The platform consists of applications designed for different steps of data processing in the clinical trial study, such as CRF Annotation, Generate SDTM Spec, SDTM Auto-Mapping, Batch Run, etc.

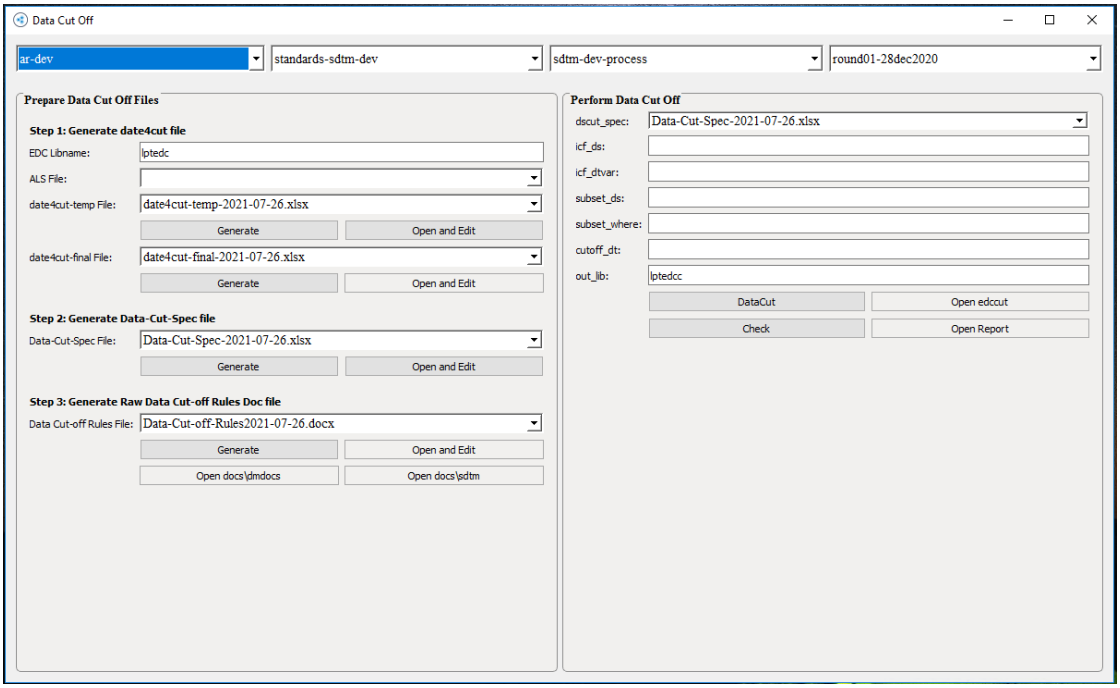

Figure 8 Dizal-Integrated Statistical Computing Platform-Data Cut-Off

CONCLUSION

Usually, data cut is performed on single RAW data set or on individual SDTM domain. Either approach requires writing simple code repeatedly, which is time consuming and easy to make mistake. For this problem, many articles have provided different, mostly theoretical, methodology. This paper presented a practical solution to centralize the data cut-off, store all relating data during the process (raw data, date kept, data removed), and generate study specific instruction files (DCO-Specification.xlsx and DCO rules.docx). The instruction files could help other teams to better understand the process and enhance cross-function communication. For programmers, the workflow would reduce manual work, increase data traceability, and make quality control more convenient. By using the iSCP, users could simply follow each step listed in the left panel to prepare all the essential files, then go to the right panel to apply DCO macro on RAW data and validate the results with the CHECK option, which further simplify the DCO and improve efficiency.

REFERENCE

[1] Ann Croft. “Implementation of Data Cut Off in Analysis of Clinical Trials”. 2018. Proceedings of PharmaSUG 2018 Conference. https://www.pharmasug.org/proceedings/2018/DS/PharmaSUG-2018-DS19.pdf

[2] Anthony L. Feliu, Stephen W. Lyons. “Leveraging SDTM Standards to Cut Datasets at Any Visit”. 2013. Proceedings of PharmaSUG 2013 Conference. https://www.pharmasug.org/proceedings/2013/DS/PharmaSUG-2013-DS02.pdf

[3] Lewis Meares. “Can You Cut It? Implementing the Data Cut-off”. 2018. Proceedings of PharmaSUG 2018 Conference. https://www.lexjansen.com/phuse/2018/dh/DH06.pdf

ACKNOWLEDGE

The author would like to thank Dizal Statistical Programming team for their review and advice on this paper.

CONTACT INFORMATION

Shang Shi
Tel: +86 15010118184
Email: Shang.Shi@dizalpharma.com

Weibin Cai
Tel: +86 15201507208
Email: Weibin.Cai@dizalpharma.com

Zhiping Yan
Tel: +86 13691423469
Email: Zhiping.Yan@dizalpharma.com

Dizal Pharmaceutical Co., Ltd
Address: Room 2106, Tower 2, China Central Place (CCP) Office Building, NO.79 Jianguo Road, Chaoyang District, Beijing, China
Zipcode: 100025
www.dizalparma.com