PharmaSUG 2015 - Poster PO14

Subset without Upsets: Concepts and Techniques to Subset SDTM Data

Jhelum Naik, PPD, Wilmington, NC
Sajeet Pavate, PPD, Wilmington, NC

ABSTRACT

For certain studies, the critical endpoints may occur at an interim visit rather than the end of the study OR the safety data may need to be monitored at periodic intervals. These trial designs dictate the requirement for subsetting of CDISC SDTM (Study Data Tabulation Model) data for an Interim analysis to generate CDISC ADaM (Analysis Data Model) which supports this analysis.

We demonstrate via case studies and operational details that it is more efficient to subset the SDTM data vs. the raw data. We explain three approaches to subset the SDTM data: Hybrid, Visit and Calendar Date. The hybrid approach looks at a combination of Visit (unique subset date per subject) and a common “calendar” cut-off date (same date for all subjects) to subset domains. The Visit based approach hinges upon identification and use of a unique “subset” date per subject based on the subject’s visit date to subset the data across all SDTM domains. This unique date can be generated based upon the status of a subject in the study. The Calendar Date approach depends upon a fixed calendar date which is the same date for all subjects. In this paper, we highlight the differences between the three approaches. We show how the data can be further cut to keep only a subset of subjects in the final data. We provide operational details on the subset technique, issues to watch out for in certain key domains such as SE, EX, AE etc., Quality Checks to ensure data consistency across SDTM domains, issues in OpenCDISC and handling missing and/or partial dates.

INTRODUCTION

As per the FDA Guidance document released in December 2014, submissions in electronic format will need to be submitted in CDISC SDTM compliant format. In most studies, SDTM is submitted at the end of the trial when all the data is collected. However, in some studies due to the unique nature of the trial design the SDTM data may need to be subset for the purposes of submission/analysis during a certain interim time point of the study.

Two examples of trial designs that may require data to be subset are listed below.

1. The primary and secondary endpoints of the study occur at an earlier time point than the actual completion of the study and may necessitate the need for an earlier submission rather than hold off until the completion of the study (e.g. a Week 24 analysis in an ongoing 5 year study).
2. There may be a need to subset data based on a calendar date for annual/periodic safety monitoring of the study (e.g. a yearly Development Safety Update Report (DSUR) safety review).

The challenge at hand is to subset clinical trial data at certain point although subjects are at different time points of the clinical trial. In SDTM compliant format, this subset of data should contain all the required information and not have any data points beyond date/period of interest.

This paper provides all the concepts and techniques required to subset the SDTM data and the operational details required to make this a successful endeavor. Since the SDTM structure is consistent across protocols and therapeutic areas, designing a system to subset SDTM can lead to time and budget efficiencies.

WHERE TO SUBSET THE DATA

One of the most common approaches would be to subset the data at the raw data level. This may seem easier and intuitive in theory. However, we do not recommend this approach for the following reasons:

- Raw data can be inconsistent across datasets as raw data may come from Electronic Data Capture (EDC) or from external vendors such as Central Laboratory.
- Since SDTM is standard, it is easier and simpler to implement the subset algorithm on a similar “sister” study even though the collection Case Report Forms are not exactly same between the two studies.

We recommend to subset the data when the full SDTM data is ready but prior to the creation of the Supplemental data sets. One reason is that all the transformation from raw to SDTM variables as well as any derivations is completed at this stage, and hence it provides us with a consistent platform to apply the subset algorithm across multiple domains. This also ensures that when supplemental data sets are created after the subset, the record in the supplemental data set correctly relates back to the record in the parent data set.

TYPES OF SUBSETS DISCUSSED IN THE PAPER

HYBRID APPROACH

The hybrid approach looks at a combination of Visit (unique subset date per subject) and a common “calendar” cut-off date (same date for all subjects) to subset domains.

In the trial design of a study, a particular scheduled visit may be of utmost importance. The date of this visit will be different for subjects depending on when they enrolled in the study. In this paper this visit will be referred to as the Subset Visit.
The calendar date is a fixed date which is typically chosen to be a date that is beyond the last enrolled subject’s Subset Visit on that study. This date will be the same for all subjects. The number of visits included in the subset will depend on how early or late the subject was enrolled in the study. Some subjects may have more visits as compared to others if they were among the first to be enrolled to the study.

The study team will identify which domains will be subset based on visit date or on common calendar date. A MS Excel spreadsheet can be created with this information. Please note that there are certain domains that typically will not be subset and will keep all the data. Examples of these domains are DM and MH (which typically do not have any change in data beyond screening).
This spreadsheet can be read in by the Subset program to determine which algorithm needs to be implemented on the domain. For example, the domains having a subset type of ‘VIS’ will be cut based on the visit date, while the domains with subset type of ‘CAL’ will be cut based on the calendar date and the domains with the subset type of ‘ALL’ will not be subset.

The subset program will need to identify which subject’s data will be cut. In general, there will be 4 kinds of subjects in a study.

- Subject with STATUS=’SCF’. These subjects are screen failures. The data for these subjects will not be subset.
- Subject with STATUS=’ERT’. These subjects are Early Terminated subjects that have discontinued Treatment prior to the Subset Visit. The data for these subjects will not be subset.
- Subjects with STATUS=’ONG’. These subjects are ongoing subjects that have not discontinued treatment but have not yet reached the Subset Visit or subset date. The data for these subjects will not be subset.
- Subjects with the STATUS=’CUT’. These subjects have completed the Subset Visit or have data beyond the subset date. The data for this subject will be subset.

The subject’s status can be determined from the following domains: Demographics (DM), Disposition (DS), Subject Visits (SV), Subject Elements (SE) and any other domains that contain relevant information to determine if they reached the Subset Visit.

A permanent SAS data set called ‘SUBS’ can be created that stores the following information: STUDYID, USUBJID, Subject Status and the Subset Date which is the Subset Visit Date. The calendar subset date can either be stored in this data set or can be stored in a global macro variable (see Operational Details below).
This data set will be used in the programs to determine the status of the subject and the subset date.

For some teams, that have individual programs for each SDTM domain, they may like to implement the algorithm to subset the data in each of these programs.
Our paper recommends creating a separate program if possible as we can create a macro for the algorithm that can be called to process each of the domains. It may thus be easier to maintain through the life of the study.

As noted above, the subset algorithm will be run on the FULL data. Ideally, the supplemental data sets have not been created yet.

The date associated with the event/assessment will be compared with the subset date (either Subset Visit date or calendar date) to determine if a record will be subset or not.

In general, the dates in the Findings class of domains will have dates in the format –DTC while in the rest of the classes the dates will be in the format –STDTC. Note we will compare the start dates with the subset dates.
For each domain, note the date that will be used for the comparison with the subset date. This date will be converted into a numeric date for comparison purposes.

If a domain needs to be restricted by Subset Visit, then the data is subset as follows:

- For domains that collect visit information, if the visit is a scheduled visit and if the visit number is less than or equal to the Subset Visit number then keep that record in the subset.
  A Quality Check can be included to check the event/assessment date against the subset date for the subject to ensure the correct record has been maintained in the subset.
- For domains that collect visit information, if the visit is an Unscheduled visit then compare the event/assessment date with the visit subset date stored in the data set 'SUBS'. If the event/assessment date is less than or equal to subset date then it will be maintained in the subset. Note that some studies may add a buffer of certain number of days to ensure that information collected few days beyond the scheduled visit is kept in the subset data.
- For domains that do not collect visit information, compare the event/assessment date with the visit subset date stored in the data set 'SUBS'. If the event/assessment date is less than or equal to subset date then it will be maintained in the subset.

If a domain needs to be subset by Calendar date, then the data is subset as follows:

- If the event/assessment date or visit date is less than or equal to the calendar subset date then that record will be maintained in the subset.

The order of the data sets being subset is important as there may be domains that will be dependent on others. For example, Subject Visits (SV), Disposition (DS) and Subject Elements (SE) will need to be subset prior to other domains.

For domains that are dependent on these domains, some programming code that creates the full SDTM compliant domain will need to be repeated in the subset program. This code will need to be repeated in order to use the new SV\DS\SE data sets that have been subset. Please see 'Special Domains and Considerations' section for additional details.

VISIT BASED APPROACH

The Visit based approach depends upon identification and use of a unique “subset” date per subject based on the subject's visit date to subset the data across all SDTM domains. This unique date can be generated based upon the status of a subject in the study and will be differ among subjects.

The Visit Based approach follows similar steps as the Hybrid approach. In this approach domains are subset based on the Subset Visit of interest. The Subset Visit date for the subject is considered the subset date for all domains that need to be subset. There will still be some domains such as DM and MH that will not be subset similar to the examples given in Hybrid approach.

We still need to identify the four different types of subjects in the study. All the domains that need to be subset will have subset type of 'VIS' and the rest will have subset type of 'ALL'.

The rest of the steps as applicable for the Subset Visit Date cut from the Hybrid approach will be followed.

CALENDAR DATE APPROACH

The Calendar Date approach depends upon a fixed calendar date which is the same date for all subjects. The subjects that have enrolled earlier in the study will have more visits and data as compared to the subjects that have enrolled later.

The Calendar Date approach follows similar steps as the Hybrid approach. In this approach, domains are subset by comparing the date of event/assessment with the fixed calendar date. There will still be some domains such as DM and MH that will not be cut similar to the examples given in Hybrid approach.

We still need to identify the four different types of subjects in the study. All the domains that need to be subset will have subset type of 'CAL' and the rest will have subset type of 'ALL'.

The rest of the steps as applicable for the Calendar Date subset from the Hybrid approach will be followed.

OPERATIONAL DETAILS

The algorithm described for the above three approaches is robust enough to be applied to most clinical trials. In order to reap the maximum benefits, it is essential to have a well-tailored plan that should involve functional groups other than Programming, such as Clinical Management and Data Management. Each team member should be aware of which data points require review, cleaning and will be included in the subset.
Documentation should be in place that discusses how the data points will be "locked" in the Electronic Data Capture (EDC) system, which data points could potentially require an update after the "lock" and how to handle such data changes. Additionally, data entry strategies should be devised to ensure all required data has been entered and to have a "black-out" period when data entry should be put on hold to avoid new data entry as we are performing these subsets on ongoing trials.

From a technical programming perspective, the following steps can add value during the execution and quality validation of the subset data.

1. Create a global subset flag macro variable to distinguish subset runs from Full SDTM runs. If needed, an alert can be added to the SAS log programs to denote the program is being run in subset mode.
   This global macro variable can also be utilized to run different sections of the programming code depending on whether it is the full run or the subset run.
2. Create a global calendar subset date macro variable to populate the calendar subset date. This macro variable can be used consistently across all SDTM domains in the hybrid and Calendar Date approach.
3. Assign a status to each subject, as follows:
   a. SCF – Screen Failure Subject; all data is included
   b. ONG – Ongoing Subject; Subject has not yet reached subset visit or date, did not discontinue prior to the subset visit; all data is included
   c. CUT – Subject has reached or gone beyond the subset visit or date and needs to be subset;
   d. ERT – Subjects has Early Terminated prior to the subset and all data is included.

By including the status, we can quickly identify what data will be included for a given subject and respond promptly to any data management and/or clinical questions.

SUBSET OF SUBJECTS

There may be instances where the subset of data may be required only for some key subjects. In these cases, we recommend creating a data set called 'SUBPT' from the Demographics domain containing the list of key subjects that satisfy the criteria as decided by the study team. At the beginning of the subset process, we recommend that this data set be merged to all data sets. Only subjects present in the SUBPT dataset will be kept in the corresponding data set. This data set with only the key subjects will then pass through the rest of the subset algorithm.

CASE STUDIES FOR EXAMPLES OF THESE APPROACHES

The authors of this paper had to devise a strategy to subset the SDTM data for a five year study. The primary efficacy endpoint for this study is six months after baseline. It was critical to subset the data at the six month visit as the plan was to submit the data for regulatory approval (along with other trials where the data was subset in a similar manner). Note that the subjects and data entry continued for beyond the six month visit.

In addition, we also had to adapt this Visit based approach to the calendar date subset approach for an integrated safety submission. By working closely with our Clinical and Data Management counterparts, we were able to generate high quality subset SDTM data transfers and helped to secure a regulatory approval for the investigational product.

A few months after receiving regulatory approval, we had to subset the SDTM data on the same trials using the Hybrid approach to support paper publications.

SPECIAL DOMAINS AND CIRCUMSTANCES

Since some domains are dependent on other domains in SDTM, we need to consider special circumstances while implementing the subset algorithm.

DEMOGRAPHICS (DM)

This domain will not be subset and all subjects data will be included in the final data set. There are some variables in the DM domain that will need to be updated to ensure that this domain does not contain any information beyond the subset date. The variables RFENDTC, RFXENDTC, RFPENDTC, if populated, will need to be updated. These variables generally contain the Date/Time of Last Treatment Exposure or End of Participation for a variable. Since the subset algorithm is run on full data, there is a possibility for some subjects that these date/time variables will be populated with a date beyond the subset date. In such cases these variables will need to be updated as either NULL or to the last available exposure date prior to or on the subset date.

These variables in DM may also be used in other domains for calculations in some variables. For example, the timing variables in certain domains will need to be re-calculated using the updated variables in the DM data set. For example, --ENRTPT variables in domains such as AE and CM will need to be updated.

SUBJECT ELEMENTS (SE)

This domain will require the most changes as compared to the other domains. This domain is created from DM, DS, EX and other domains. It is recommended to utilize the previously subset domains for the creation of the SE domain. In some studies, SE domain may need to be re-created in the subset program rather than subset the full SE domain. Depending on study design, the study team can decide the best course of action

EPOCH VARIABLE

The EPOCH variable, if populated, in certain domains, such as EX and DS for example, will need to be verified and may have to be re-generated to ensure that it lines up correctly with the Subset SE dataset.

QUALITY CHECKS (QC)

We recommend creating checks that compare and check data across domains for consistency. There should be checks that compare data between SE and DS; DS and EX for consistency. For example, if the Last Study Dose is collected on the CRF and is stored in DS then it should be consistent with the available subset data in the EX domain.

For a Visit based approach, there should be a check that lists the maximum visit per subject to ensure that it is not beyond the Subset visit. Similarly for the Hybrid and the Calendar approach, the maximum visit date and/or event/assessment date should not be beyond the subset date.
There should also be a check to ensure that all visits in the domains are represented in SV and vice-versa.

MISSING EVENT/ASSESSMENT DATES

There may be scenarios where an assessment is not done at a given visit. For such records, the assessment date will be missing but the visit date may be populated. In such scenarios consider the visit date in the comparison. If the visit date is also missing then check if the visit is present in the subset SV domain and keep the record accordingly.

If event dates are missing for domains where visit information is not collected then include a quality check to output such records for queries. These records should be kept in the subset as we cannot determine if the date of the event occurred after the subset date.

PARTIAL EVENT/ASSESSMENT DATES

The study team will need to discuss on how to handle partial dates in the subset algorithm. Our recommendation is to be conservative and keep such records.
For example, if year is present but month and day are missing then do as follows:

- If the year is prior or equal to the year of the subset date then include the record else delete

If Year and Month are present but day is missing then do as follows:

- If the month and year is prior or same as the year and month of the subset date then include the record or else delete.

VALIDATION CHECKS

While reviewing validation checks such as OpenCDISC validator report, keep a look for false positives. For example, data in Disposition (DS) may not have a completion record as subjects are still ongoing at the time of the Subset.

CONCLUSION

In the current landscape where regulatory agencies prefer that data should be submitted using CDISC SDTM standards, it makes good business sense to harness the consistency in SDTM to subset the data. Using a simple framework, we have demonstrated how the SDTM data can be subset in three different ways as per business requirements. This framework is not dependent upon any protocol specific information and it can be implemented with ease across multiple therapeutic areas.

Based on our experience of successful implementation of these techniques, we have provided key items for consideration and common challenges. We hope that this may benefit readers so that they can subset data without being “upset”.

REFERENCES

Providing Regulatory Submissions In Electronic Format — Standardized Study Data. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334.pdf

CDISC SDTM Standards information. Available at http://www.cdisc.org/sdtm

ACKNOWLEDGMENTS

The authors would like to thank Ken Borowiak for his thorough review and other PPD colleagues for their support. They would also like to thank Brandon Graham for his guidance and review.

RECOMMENDED READING

Feliu, Anthony; Lyons, Stephen. 2013. “Leveraging SDTM Standards to Cut Datasets at Any Visit”. PharmaSUG 2013. Available at http://www.pharmasug.org/proceedings/2013/DS/PharmaSUG-2013-DS02.pdf

DISCLAIMER

The content of this paper are the works of the authors and do not necessarily represent the opinions, recommendations, or practices of PPD.

CONTACT INFORMATION

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

Jhelum Naik
PPD
929 North Front Street
Wilmington, NC 28401
Work Phone: +1 910 202 4761
E-mail: jhelum.naik@ppdi.com

Sajeet Pavate
PPD
929 North Front Street
Wilmington, NC 28401
Work Phone: +1 910 558 8622
E-mail: sajeet.pavate@ppdi.com