PharmaSUG 2017 - Paper BB14

A SAS®sy Study of eDiary Data

Amie Bissonett, inVentiv Health Clinical, Minneapolis, MN

ABSTRACT

Many sponsors are using electronic diaries (eDiaries) to allow subjects to enter study data themselves, such as daily events, concomitant medications taken, and symptoms that occur. Depending on the study, subjects may enter data at varying time increments, from weekly or monthly history up to a daily account of activities and events. The timeliness of the data entry as well as the cleanliness of the data make a big impact on deriving SDTM and ADaM data sets and how the analysis will be performed. This paper goes through different scenarios and gives some tips to help from data cleaning, setting up the variable derivations, and programming the analysis data sets.

INTRODUCTION

eDiary data can come in many forms and be used to collect multiple types of data, all of which can have their own challenges. The raw datasets, likely from an external vendor, need to be cleaned, annotated to the SDTM domains required, and finally transformed in to the analysis ready ADaM data sets.

The format of the raw eDiary data is similar to other eCRF data. Each screen on the eDiary device a subject fills in the fields very similar to a site filling out an eCRF. The layout of the raw data must be thoroughly reviewed and annotated before any programming can begin, and the study's analysis endpoints will impact how the raw data is transformed.

- What kind of data is the subject entering? Safety? Efficacy?
- What is the timeliness of the data entry? Daily/weekly/monthly? Multiple entries per day?
- Is the data entered based on the time increment or is it event based?
- What is the detail of the raw data?
    - One record per time increment (day will be used going forward for examples)
    - One record per event
    - Multiple records per event if the event spans more than one day
- Are the data sets from different screens connected by identifier variables, i.e. an ID variable?
- Can the subject enter data retroactively?

This paper is based on an events based study for migraine headaches where the subjects enter data daily as well as for all headache events and medications taken, which both can occur multiple times per day.

DATA CHECKS

eDiary vendors should have edit checks to clean the data prior to sending it to the sponsor, however, they are generally more basic checks and could miss data issues that affect the data analysis. Subjects aren't as well trained as site staff so it is inevitable that data issues will arise. Some examples of data issues that the vendor might not catch follow.

- Duplicates – while the vendor likely checks for certain duplicates, the analysis can determine additional combination of variables that cannot be duplicated.
    - Daily symptoms that should only be done once per day are entered more than once. More than one headache can be entered per day so this occurs frequently.
    - Headaches entered using the same ID number more than once

  - Medications entered more than once with the same dose and time taken
- Event start and end date/time issues
  - Headache, aura or sleep times that start between the start and end of another headache
  - Headache, aura or sleep end times occurring prior to the start time
  - More than one entry with a headache end date/time for the same headache ID.
  - A headache entered, which requires a start time, without an end time.

Once you find data issues with the data checks, where does that information go? Each check can have an output data set containing all of the data with that check's issue. The study team needs to determine which of the issues are important and need to be reported to the vendor to be fixed. An excel spreadsheet makes an easy layout to keep track of each check, and a new column to enter the number of issues for each check each time the checks are run and programmer comments to the reviewers on what to review.

Some SAS®sy SQL procedure code to summarize the check information makes the check process quick and easy. This gives a quick count of the total number of records found in a check and the number of subjects affected:

```
proc sql ;
  create table ckdlydup as
    select distinct count(*) as nrecs, count(distinct subjid) as nsubj
      from dly_dup
  ;
quit ;
```

The data check data sets should be set up to list the information needed to send to the vendor so they can easily identify which records are affected:

```
proc print data=ha_dup ;
  var subjid hednum1n drydt hastdtm harendtm slpstdtm slpendtm ;
run ;
```

HOW TO DEAL WITH DATA THAT DOESN'T GET CLEANED

It needs to be determined within the study team if certain data issues can remain in the data. Data that is circumstantial and not used in analysis may not be worth the time and resources to fix the issues. It may be sufficient to select the distinct values of certain fields, i.e. medication duplicates, if it is only informative that the subject took any of the specific medication on that date.

SDTM DATA SET CREATION

Depending on the project, different SDTM domains will be required. This paper focuses on the following domains:

- CC – Clinical Classification, an interventions domain used for migraine specific concomitant medications
- CE – Clinical Events, an events domain detailing specific headache, aura, and sleep information, such as start and end date/times.
- DF – Disorder Findings, a findings domain which includes additional information on the signs and symptoms of the headaches

There were three raw data sets containing the eDiary data, daily occurrences, headache events, and medications. All raw data sets contain a VISITDT variable, which is the date/time stamp of when the subject entered the data in the eDiary device, and a DRYDT, which is the diary date that the subject specifies that they are entering data for. The diary date is the basis for deriving the primary and secondary endpoints which are whether a subject has a migraine or headache day based on criteria using headache duration and associated symptoms and occurences.

CE and DF domains contain all of the information for all headaches, auras, and sleep. There were two raw data sets used for both of these SDTM data sets, both in horizontal format. One contained headache specific information (headache raw), including the start and end times, sleep start and end times associated with a specific headache, the type of headache, migraine versus non-migraine, and headache specific symptoms, such as severity, pulsating, nausea, etc. The diary date was not useful here because headaches spanning more than one day had multiple diary dates associated with a specific headache. The second contained overall daily occurrences (daily raw), such as migraine interference with daily activities, missing work or school, and whether any auras occurred.

The CE and DF domains are an event and finding domain, respectively, and as such, the data in both of the raw data sets needed to be split across the two SDTM data sets. The headache raw data contained a headache ID value that allows for easy identification of headache information between CE and DF. The daily raw data is not associated with an ID variable. The raw data also needed to be transformed in to vertical format to conform to the SDTM standard format.

Additional complications arose because it was determined to collapse all records for a headache ID in to one record in CE. The headache raw data had one record per date for headaches that spanned more than one day. Any sleep times associated with a headache ID were then output to a separate record in CE. Aura start and end times were also added from the daily raw data, but were not associated with a headache ID.

The DF data set takes all of the symptoms from headache raw data and the daily occurrences from daily raw data and transposes them to one record per symptom or occurrence per day. Similar to CE, the headache symptoms in DF were associated with a headache ID, but the daily occurrences were not.

ADAM DATA SET CREATION

The data cited in this paper requires multiple transformations from raw data to SDTM to ADaM. The raw data is in horizontal format for all data sets. The SDTM data sets are transposed to vertical format. Finally, SDTM data sets and transformed back to horizontal format and merged together prior to deriving the analysis ready variables. In addition, the CE domain compressed headaches from multiple records when a headache spanned multiple days down to one record and ADaM data required one record per day for all headaches.

As mentioned previously, the basis for the analysis was to determine if a subject had a qualified headache (non-migraine) or migraine each day that they were in the study, so a record is required for every day. The collapsed CE headache information needed to be expanded back out to one record per day for multi-day headaches.

The primary/secondary efficacy analysis was a monthly summary of the change in the number of migraine/headache days. The ADaM data set used for analysis was a Basic Data Structure (BDS) monthly summary of all primary and secondary endpoints as the individual parameters.

Three interim ADaM data sets were used to derive the final monthly summary statistics, two of the more complex of which will be discussed.

A daily headache data set, ADHA, was first derived from SDTM. Since the diary date was not kept in CE for the headache records it was determined to set the headache start date as the analysis date (ADT) and use that for merging with other data as needed. Where available in other data, ADT was set to the diary date. This data set only kept records for dates where a headache occurred.

When a headache spans more than one day, records are derived for each date and the end and start times need to be set for the end and start of each day. This code excerpt from a data step loops through from the headache start date to the end date and assigns ADT to the particular date of the headache, or headache start date as mentioned previously. When the headache spans midnight, the headache end date/time for that date needs to be assigned to midnight and the headache start date/time for the following day is also set to the same time starting the next calendar day. The following DATA step code demonstrates derivation of the daily records:

```
do todayst = hastdt to haendt ;
  adt = todayst ;
  /* prior to this loop, hasttmp = raw data headache start datetime */
  hastdtm = hasttmp ;

  /* haentmp = raw data headache end datetime */
  if todayst = haendt then do ;
    haendtm = haentmp ;
  end ;
  else do ;
    haendtm = dhms(todayst, 24, 0, 0) ;  /* midnight of current day */
  end ;

  /* output record for calendar day ha */
  output ;

  /* set value for hastdtm for next day */
  hasttmp = haendtm ;
end ;  /* do hastdt to haendt */
```

The complex issue for this data set was expanding the CE data from the collapsed one record per headache ID back to one record per headache ID per day and then merging the daily occurrences without a headache ID, and merging auras and medications based on a specified date/time range merge. Auras are only associated with a headache when they overlap with a headache or if the aura occurred within an hour of the headache start time. A range merge is required for this, so PROC SQL is the way to go:

```
proc sql ;
  create table haaura as
    select *
      from (
        select distinct a.*, austdtm, auendtm
          from (
            select *
              from allha
          ) a
          LEFT JOIN
          aura b
          on a.usubjid=b.usubjid &
            ((. < hastdtm <= austdtm <= haendtm) or
             (. < hastdtm <= auendtm <= haendtm) or
             (. < austdtm <= hastdtm < auendtm) or
             (0 <= (hastdtm - auendtm) / 60 <= 60)
            )
      )
      group by usubjid, haid, adt
      order by usubjid, haid, adt, austdtm
  ;
quit ;
```

The second interim ADaM data set, ADDA, expanded upon ADHA and used a shell to fill in all days that a subject was on the study, from their first to last eDiary entry dates. Information included here that was excluded from ADHA was any auras or medications that didn’t meet the merge criteria, based on headache and aura or medication start and end times, and any daily occurrences entered on a date without a headache. This information not associated specifically with a headache could still determine certain endpoint flags. ADDA contained several Yes/No flags for each day, including whether that day the subject entered any eDiary data, was it a migraine day, a headache day, did they use migraine specific medications other than the investigational product, as well as the total duration of migraine and headache hours for that day.

Finally, the monthly summary data set, ADMO, is derived from a combination of the three interim data sets. While ADHA was used to derive specific daily flags and daily headache duration in ADDA, the symptoms and severities of them in ADHA were not kept in ADDA and were required for the final derivations. In addition to summing the information for each month in the study, a proration was applied to account for subjects not completing their eDiary every day. All of the derived endpoints were then transposed to the BDS format and each assigned their given PARAM value.

WHAT I WISH I KNEW BEFORE PROGRAMMING WITH EDIARY DATA

eDiary specifications need to be thoroughly reviewed before and during programming with the data. Additionally, the specifications need to be just that, specific. Subjects entering data, especially those having a migraine headache, can make countless errors, repeat their entries, or miss entering data. It is imperative to have set rules on how to handle all of the complexities when working with this type of data.

While the vendor in charge of the eDiary devices and raw data sets should be doing data checks, it is incredibly helpful in learning the data and the issues common with it to do a set of data checks prior to converting the raw data to SDTM data sets.

Be prepared for changes. The study detailed in this paper was for a new project so there were a lot of questions asked and changes made throughout the study programming and analysis, even in to post-hoc analysis. Writing code as robust as possible, including using macros to automate similar code, while trying to foresee possible changes will save much time and energy. Comment, comment, comment. Keep track of all of the changes made throughout the study so that there is a trail of what was and what is.

CONCLUSION

Data entered by study subjects is likely to be ridden with data issues. They don’t have the training that site staff do, and even data entered at study sites encounter many errors. Programming with eDiary data isn’t all that different than any other data, but the complexity of the data can certainly be in the upper echelon. Take the time to learn the specifics of the raw data, the requirements of the SDTM domains, and how the final ADaM analysis parameter values need to be derived.

RECOMMENDED READING

- *CDISC www.cdisc.org*

CONTACT INFORMATION

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

Amie Bissonett
inVentiv Health Clinical
amie.bissonett@inventivhealth.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.