PharmaSUG 2015 – Paper HA01

Statistical Analyses Across Overlapping Time Intervals Based on Person-Years

John Reilly, Dataceutics, Inc., Pottstown, PA
John R. Gerlach, Dataceutics, Inc., Pottstown, PA

ABSTRACT

Consider an ongoing observational study consisting of subjects who take a drug to treat a disease that causes serious health problems, such as renal failure, or even death. A scheduled analysis produces a report showing several items of interest (e.g. incidence rate) across *overlapping* time intervals that are based on person-years, that is, a subject's exposure to the therapy treatment. Consequently, a subject may contribute to multiple time intervals. Although a subject might have more than one event only the earliest event is used. Conversely, a subject might not have any events, yet contributes to the overall person-years for appropriate time intervals. This paper explains an intuitive method for augmenting the analysis data set so that the overlapping time intervals are represented, accordingly.

INTRODUCTION

Even after a clinical trial and regulatory approval, long-term observational studies are often implemented to monitor the safety and efficacy of a drug. Numerous reports are generated ranging from descriptive statistics to statistical modelling for outcomes research. An analysis might have a longitudinal component indicating some item of interest across several time intervals (e.g. 2-3 months). However, *what if the time intervals were not mutually exclusive?*

Consider an analysis based on person-years of exposure to a drug. Now, consider if the time intervals overlap, for example: exposure for more than one year and up to three years; and, exposure for more than two years and up to five years. These two time intervals overlap. Thus, for example, given 2.9 person-years, a subject would contribute 1.9 person-years to both intervals, since both intervals begin after one year on therapy (i.e. 2.9 minus 1.0 year). Before listing the overlapping time intervals of interest, let's review Interval Notation, as shown below.

$(a,b) \rightarrow \{ x \in R \mid a < x < b \}$ $[a,b) \rightarrow \{ x \in R \mid a \le x < b \}$
$(a,b] \rightarrow \{ x \in R \mid a < x \le b \}$ $[a,b] \rightarrow \{ x \in R \mid a \le x \le b \}$

Below is a list of 14 time intervals, several of which overlap (e.g. intervals 7 and 8). Notice that Interval #1 is *closed*, whereas, the others are *open*; that is, the lower limit value is **not** included. Thus, for example, a subject having 1 person-year would contribute to intervals 1 through 4, but not interval 5, since it is open on the lower limit. Although the person-years determines the intervals, it is the interval itself that determines the *apportioned* number of years. In this example, interval 1 would receive 1.0 year while the other three intervals would receive half that.

Interval	Notation	Description	
1	[0, 0.5]	0 to 0.5 year	
2	(0.5, 1.0]	Greater than 0.5,	up to 1 year
3	(0.5, 5]	"	up to 5 years
4	(0, 1]	Greater than 0,	up to 1 year
5	(1, 2]	Greater than 1,	up to 2 years
6	(1, 3]	"	up to 3 years
7	(1, 5]	"	up to 5 years
8	(2, 3]	Greater than 2,	up to 3 years
9	(2, 5]	"	up to 5 years
10	(3, 4]	Greater than 3,	up to 4 years
11	(3, 5]	"	up to 5 years
12	(4, 5]	Greater than 4,	up to 5 years
13	(5, α]	Greater than 5 years	
14	(6, α]	Greater than 6 years	

Table 1. Time intervals.

THE ANALYSIS DATA SET

Even though the observational study is not intended for submission to a regulatory agency, because of its popular use, the CDISC naming convention is used for the sake of clarity. However, it should be noted that the ADaM Time-to-Event (ADTTE) data set is not appropriate for this analysis. The analysis variable (AVAL) in ADTTE is parameter specific, such as the time (days) to an event; whereas, the Person-years discussed in this paper is a fixed attribute of a subject (i.e. the number of years on therapy). The intended analysis answers the question, for example: *How many subjects had a cardiac arrest while on therapy for more than 1 year and up to 3 years?*

The data is extracted from a longitudinal database for which there are two data sets of interest:

- **ADSL – Subject Level**
  - USUBJID — Unique Subject Identifier
  - SEX — Gender
  - AGE — Age (years)
  - TRTSDT — Date of First Exposure to Treatment
  - TRTEDT — Date of Last Exposure to Treatment

- **ADEVT – Events**
  - USUBJID — Unique Subject Identifier
  - EVTDT — Event Date
  - EVTANY — Any Event
  - EVTCAT — Event Category
    - 1 Cardiac Arrest
    - 2 Ischemic Stroke
    - 3 Renal Failure
    - 4 Death

The first Data step below defines the cohort based on selection criteria, but also computes the time of exposure to therapy in person-years. For this analysis, the earliest event is of interest, given that there is only one event on a given date. The SORT and subsequent Data step selects the earliest events of interest. Then, the final Data step merges the ADSL and ADEVT02 data sets keeping all subjects regardless of whether a subject had an event.

```
data adsl;
   set ad.adsl;
   pyears = (trtedt - trtsdt) / 365.25;
   if < selection criteria defining cohort > ;
   keep usubjid sex age trtsdt trtedt pyears;
run;

proc sort data=ad.adevt(keep=usubjid evtdt evtany evtcat) out=adevt01;
   by usubjid evtdt;
   where evtcat in(1,2,3,4);
run;

data adevt02;
   set adevt01;
      by usubjid;
   if first.usubjid then output;
   drop evtdt;
run;

data adset01;
   merge adsl(in=pop) adevt02;
      by usubjid;
   if pop;
run;
```

The unit of analysis of the data set ADSET01 is: One record per subject who *might* have had an event (e.g. cardiac arrest) while on therapy. At this point, the data set is inadequate for the intended analysis since there is no

categorical variable denoting the overlapping time intervals. In fact, the continuous variable PYEARS is used to create two new variables: a categorical variable denoting the time interval (INTERVAL) and a continuous variable denoting the person-years apportioned to that same interval (PYRS). More importantly, the data set must be augmented such that there is an observation for each time interval as a function of a subject's person-years. For example, as noted earlier, if a subject has 2.9 person years of treatment, then the process generates eight observations and distributes the person-years, accordingly.

Table 2 shows a partial listing of the initial data set prior to augmentation. At first the variable EVTANY seems redundant, even superfluous, next to EVTCAT (Event Category), but it's not. However, these Event variables do pose a problem (a caveat in the data processing) that will be addressed later.

USUBJID	SEX	AGE	TRTSDT	TRTEDT	PYEARS	EVTANY	EVTCAT
1001	1	40.6	26SEP2002	25JUN2009	6.75	1	3
1002	1	55.8	14SEP2001	02AUG2004	2.90	1	3
1003	1	32.6	04MAY2002	05MAY2005	3.00		
1004	1	17.0	20JAN2004	11JUN2007	3.39		
:	:	:	:	:	:	:	:

Table 2. Partial listing of initial analysis data set.

Table 3 shows the results for two subjects, both who suffered renal failure while on therapy for 6.75 and 2.9 years, respectively. Notice that Subject 1001 generated 14 observations while Subject 1002 generated only 8 observations, representing their appropriate intervals. The variable PYRS represents the number of person-years for a specific time interval based on the total number of person-years. Thus, for subject 1001, the Interval ">1--3" contains 2 person-years; whereas, the interval ">1—5" contains 4 person-years.

USUBJID	PYEARS	EVTANY	EVTCAT	INTERVAL	PYRS
1001	6.75	1	3	0-0.5	0.50
				>0.5-1	1.25
				>0.5-5	5.25
				>0-1	1.00
				>1-2	1.00
				>1-3	**2.00**
				>1-5	**4.00**
				>2-3	1.00
				>3-4	1.00
				>3-5	2.00
				>4-5	1.00
				>5-6	1.00
				>5	1.75
				>6	0.75
1002	2.90	1	3	0-0.5	0.50
				>0.5-1	0.50
				>0.5-5	2.40
				>0-1	1.00
				>1-2	1.00
				>1-3	2.00
				>1-5	1.90
				>2-3	0.90

Table 3. Listing of person-years (PYEARS) distributed according to their respective time intervals.

Table 3 indicates that the initial analysis data set will grow substantially. Hypothetically, given 100 subjects and 14 time intervals, the initial analysis data set could grow from having 100 observations to 1,400 observations.

Regardless of the number of events (e.g. cardiac arrest), each observation will contribute to one or more time intervals based on the person-years. Person-years will determine the time interval; whereas, the time-interval will determine the portion of the person-years. Also, recall the closed interval [0-1); whereas, the others are open intervals on the lower-limit side, which matters, albeit slightly. Table 4 shows the two rules for assigning the categorical variable INTERVAL and the continuous variable PYRS, written in pseudo-code. The two code segments differ only with respect to identifying the lower bound of a given interval.

- **Closed Interval**

```
if years GE lower_limit

    Assign interval category
    If years LT upper_limit
        then person-years = years – lower_limit
        else person-years = upper_limit
    Output observation
```

- **Open Interval**

```
if years GT lower_limit

    Assign interval category
    If years LT upper_limit
        then person-years = years – lower_limit
        else person-years = upper_limit
    Output observation
```

Table 4. Rules for assigning interval category and person-years.

THE REPORT

The report shell below (Table 5) illustrates several analysis items of interest across overlapping time intervals that indicate person years on therapy. The several items are defined, as follows:

- Composite Events — Number of subjects having any event
- Subjects at Risk — Number of subjects at risk
- Percent (95% CI) — Percentage of subjects at risk, with 95% confidence interval
- Person-years — Total person-years
- Incidence Rate — Rate of events (per 1000) with a 95% confidence interval
- Events — Number of subjects having a specific event (e.g. cardiac arrest)

Most of the items are frequency counts, along with a total sum for Person-years. However, the analysis includes percentages and incidence rates, along with their respective confidence intervals. Overall, the analysis is rather straight-forward. However, creating the appropriate analysis data set is another matter.

Regardless of the statistics involved, it is important to keep in mind that the analyses is done across **overlapping** time intervals. Thus, for the sake of brevity, this paper will focus on the frequency counts, specifically the events, rather than the more involved inferential statistics, such as the confidence interval of a Poisson statistic. The paramount issue concerns the augmentation of the initial analysis data set, which must be done correctly; otherwise, the statistical analysis is worthless.

	Years on Treatment		
	0-0.5	**>1-2**	**>6**
Composite Events	xxx	xxx	xxx
Subjects at Risk	xxx	xxx	xxx
Percent (95% CI)	xx.xx (xx.xx,xx.xx)	xx.xx (xx.xx,xx.xx)	xx.xx (xx.xx,xx.xx)
Person-years	xxx.xx	xxx.xx	xxx.xx
Incidence Rate Per 1000 (95% CI)	xx.xx (xx.xx,xx.xx)	xx.xx (xx.xx,xx.xx)	xx.xx (xx.xx,xx.xx)
Events			
Cardiac Arrest	xxx	xxx	xxx
Ischemia Stroke	xxx	xxx	xxx
Renal Failure	xxx	xxx	xxx
Death	xxx	xxx	xxx

Table 5. Report shell.

The program begins by creating two formats, one defining the overlapping time intervals and the other defining the events, as shown in the PROC step below.

```
proc format;
   value intvlf  1 = '0-0.5'     2 = '>0.5-1'     3 = '>0.5-5'
                 4 = '>0-1'      5 = '>1-2'       6 = '>2-3'
                 7 = '>1-3'      8 = '>3-4'       9 = '>4-5'
                10 = '>3-5'     11 = '>1-5'      12 = '>5-6'
                13 = '>5'       14 = '>6';

   value evttyp  1 = 'Cardiac Arrest'     2 = 'Ischemic Stroke'
                 3 = 'Renal Failure'      4 = 'Death';
run;
```

The following Data step merges the study population along with events of interest, if any. For each subject, the person-years of exposure to therapy determines the value of the two new variables:

- **INTERVAL** Time interval
- **PYRS** Portion of total person-years for a respective time interval.

The sub-setting IF statement in the DATA step ensures that the analysis data set represents the study population, along with their events, if any. For each subject, the augmentation process begins with a series of nested IF statements for each interval, first identifying an interval (e.g. PYEARS > 1.0), then assigning the portion of person-years associate with that interval (e.g. PYRS = PYEARS – 1.0). Notice the logical operator GE for the first (closed) interval and the GT operator for the remaining (open) intervals. Also, keep in mind that the augmentation process occurs regardless of whether a subject has an event. Finally, the unit of analysis has changed to “One record per subject per time interval (based on Person-years).”

```
data adset01;
   merge pop(in=pop) events;
      by usubjid;
   if pop;
   if pyears GE 0.0
      then do;
         interval = 1;
         if years lt 1
            then pyrs =  pyears - 0.0;
            else pyrs =  0.5;
         output;
         end;
   if pyears GT 0.5
      then do;
         interval = 2;
         if years lt 1.0
            then pyrs =  pyears - 0.5;
            else pyrs =  0.5;
         output;
         end;

      :         :         :         :

   if pyears GT 1.0
      then do;
         interval = 10;
         if pyears lt 5.0
            then pyrs =  pyears - 1.0;
            else pyrs =  4.0;
         output;
         end;

      :         :         :         :

   if pyears GT 5.0
      then do;
         interval = 13;
         pyrs =  pyears - 5.0;
         output;
         end;
   format interval intvlf. pyears pyrs 5.2;
run;
```

As noted earlier (See Table 3), a subject's person-years of exposure determines the values of INTERVAL and PYRS, as well as how many observations are generated from a single observation. In this case, given an input data set consisting of 1,200 observations, the augmentation process generated over 10,400 observations.

Using the FREQ procedure (not shown), it is time well-spent to study the distribution of overlapping time intervals in the augmented analysis data set, as shown in Table 6 below, such as:

- The frequency count of any interval cannot exceed the number of subjects (i.e. 1,200).
- The frequency count for the interval ">2-3" is less than the count for interval ">1-3".
- The intervals ">1-5" and ">5" are mutually exclusive and collectively exhaustive subsets.
- The interval ">6" is a proper subset of the interval ">5".
- Several intervals have the same frequency count.

INTERVAL	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-0.5	1200	11.46	1200	11.46
>0.5-1	1058	10.10	2258	21.56
>0.5-5	1058	10.10	3316	31.67
>0-1	1200	11.46	4516	43.13
>1-2	952	9.09	5468	52.22
>2-3	749	7.15	6217	59.37
>1-3	952	9.09	7169	68.47
>3-4	579	5.53	7748	73.99
>4-5	434	4.14	8182	78.14
>3-5	579	5.53	8761	83.67
>1-5	952	9.09	9713	92.76
>5-6	305	2.91	10018	95.67
>5	305	2.91	10323	98.59
>6	148	1.41	10471	100.00

Table 6. Distribution of the variable YRINTVL in the augmented analysis data set.

CORRECTING THE AUGMENTED DATA SET

There is something wrong with the analysis data set with respect to the Event variables EVTANY and EVTCAT. Table 7 shows these variables in juxtaposition, which offers a hint about the problem. During the augmentation process, these invariant attributes, were carried along across the several intervals as a function of the subject's apportioned person-years. However, these variables actually represent events where the subject's person-years is congruent with a time interval. For example, a subject having 2.9 person-years should have events associated with the interval ">1-3", but not the interval ">1-2". Therefore, these variables must be corrected.

EVTANY	EVTCAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
		9041	86.34	9041	86.34
1	1	524	5.00	9565	91.35
1	2	57	0.54	9622	91.89
1	3	235	2.24	9857	94.14
1	4	467	4.46	10324	98.60
1	5	147	1.40	10471	100.00

Table 7. Distribution of EVTANY (Any Event) and EVTCAT (Event Category).

Table 7 shows the EVENT variable EVTANY (Any Event) and EVTCAT (Event Category), which have inflated value. Keep in mind that the process of augmenting the analysis data set caused this caveat in the data.

The following Data step corrects the variables EVTANY and EVTCAT for those subjects who had an event. The SELECT statement identifies the year-interval (e.g. 0-0.5), then proceeds to determine whether the person-years is not bounded by that interval; whereupon, EVTANY is re-assigned a null value, which affects the variable EVTCAT accordingly. Also, notice the RETAIN statement that defines TOTPAT, which will be used to compute the "Subjects at Risk" item in the report.

```
data adset03;
   retain totpat 1;
   set adset02;
   if evntany
      then do;
         select(put(interval,intvlf.);
            when('0-0.5')   if not (0.0 le pyears le 0.5) then evtany = .;
            when('>0.5-1')  if not (0.0 lt pyears le 1.0) then evtany = .;
            when('>0-1')    if not (0.0 lt pyears le 1.0) then evtany = .;
            when('>1-2')    if not (1.0 lt pyears le 2.0) then evtany = .;
            when('>2-3')    if not (2.0 lt pyears le 3.0) then evtany = .;
            when('>1-3')    if not (1.0 lt pyears le 3.0) then evtany = .;
            when('>3-4')    if not (3.0 lt pyears le 4.0) then evtany = .;
            when('>4-5')    if not (4.0 lt pyears le 5.0) then evtany = .;
            when('>3-5')    if not (3.0 lt pyears le 5.0) then evtany = .;
            when('>1-5')    if not (1.0 lt pyears le 5.0) then evtany = .;
            when('>5-6')    if not (5.0 lt pyears le 6.0) then evtany = .;
            when('>5')      if not (      pyears gt 5.0) then evtany = .;
            when('>6')      if not (      pyears gt 6.0) then evtany = .;
            otherwise;
            end;
         end;
   if evtany eq .
      then evtcat = .;
run;
```

Table 8 contains the "*Before and After*" report illustrating the correction made to the Event variables EVTANY and EVTCAT. In summary, EVTANY and EVTCAT are invariant attributes that were propagated during the augmentation process; however, these variables must be utilized in the context of when (which time intervals) the event occurred. Consequently, the aforementioned DATA step was needed to correct these variables.

Before

USUBJID	YRINTVL	PYEARS	PYRS	EVTANY	EVTCAT
1052	>0-0.5	2.9	0.5	1	3
	>0.5-1	2.9	0.5	1	3
	>0.5-5	2.9	2.4	1	3
	>0-1	2.9	1.0	1	3
	>1-2	2.9	1.0	1	3
	>2-3	2.9	0.9	1	3
	>1-3	2.9	1.9	1	3
	>1-5	2.9	1.9	1	3

After

USUBJID	YRINTVL	PYEARS	PYRS	EVTANY	EVTCAT
1052	>0-0.5	2.9	0.5		
	>0.5-1	2.9	0.5		
	>0.5-5	2.9	2.4	1	3
	>0-1	2.9	1.0		
	>1-2	2.9	1.0		
	>2-3	2.9	0.9		
	>1-3	2.9	1.9	1	3
	>1-5	2.9	1.9	1	3

Table 8. Correction of Event variables.

THE ANALYSIS

Finally, the data set is ready for analysis. For the sake of brevity, the discussion focuses on all the report items except for the statistics with confidence intervals. Thus, there are several components of the analysis:

- **Counting the number of subjects**
  - EVTANY Composite Events (i.e. Any Event during the time interval)
  - TOTPAT Subjects at Risk
  - EVTCAT Events (e.g. Cardiac Arrest)
- **Computing total apportioned person-years**
  - PYRS Person-Years

Counting the number of subjects requires little more than several FREQ steps using the variables EVTANY, TOTPAT, and EVTCAT, across the overlapping time intervals (INTERVAL). For the first two steps, it is sufficient to use INTERVAL in the TABLES statement such that EVTANY is used in the WHERE statement; whereas, TOTPAT is unnecessary, since it will be valued '1' for every observation, anyway. The third step requires the EVTCAT and INTERVAL in the TABLES statement to obtain the counts for all four events. Table 9 contains a partial listing of the subject counts in a more readable format, albeit different from the report shell.

```
proc freq data=adset03;
   tables interval / list missing;
   where evtany is not null;
   title1 'Composite Events';
run;

proc freq data=adset03;
   tables interval;
   title1 'Subjects at Risk';
run;

proc freq data=adset03;
   tables evtcat * interval;
   where evtcat is not null;
   title1 'Events';
run;
```

			Events			
Time Interval	**Composite Events**	**Subjects at Risk**	**Cardiac Arrest**	**Ischemic Stroke**	**Renal Failure**	**Death**
0-0.5	54	1200	14	21	15	4
>0.5-1	20	1058	10	5	3	2
>0.5-5	128	1058	45	47	20	16
>0-1	74	1200	24	26	18	6
>1-2	38	952	13	14	7	4
:	:	:	:	:	:	
>1-5	108	952	35	42	17	14
>5-6	12	305	4	4	2	2
>5	24	305	8	7	4	5
>6	12	148	4	3	2	3

Table 9. Partial listing of Subject counts.

Computing the apportioned person-years requires the MEANS procedure (not shown) using the PYRS analysis variable, along with the class variable INTERVAL. Table 10 shows the final report (abridged) that includes several time intervals and report items.

	Years on Treatment		
	0-0.5	>1-2	>6
Composite Events	54	38	12
Subjects at Risk	1200	952	148
Percent (95% CI)	4.50 (3.38, 5.87)	3.99 (2.82, 5.48)	8.11 (4.19, 14.16)
Person-years	562.0	853.1	228.5
Incidence Rate Per 1000 (95% CI)	96.09 (72.18, 125.38)	44.55 (31.52, 61.14)	52.51 (27.13, 91.73)
Events			
Cardiac Arrest	14	13	4
Ischemia Stroke	21	14	3
Renal Failure	15	7	2
Death	4	3	3

Table 10. Final report.

CONCLUSION

Monitoring a drug involves various statistical analyses that often demands more intricate preparation of the data. Performing analyses across overlapping time intervals requires an augmented data set such that the person-years on therapy generates multiple observations so that the appropriate time intervals are represented, thereby increasing the number of observations dramatically. Also, the total person-years is apportioned accordingly for each time interval. By creating two new variables, denoting the time interval and the apportioned person-years, the augmented data set becomes well-suited to perform statistical analyses across overlapping time intervals.

ACKNOWLEDGMENTS

Thanks to the management at Dataceutics for their support in this endeavor.

CONTACT INFORMATION

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

Name: John Reilly & John R. Gerlach
Enterprise: Dataceutics, Inc.
Work Phone: 610-970-2333
E-mail: reillyj@dataceutics.com