PharmaSUG 2016 - Poster PO04

Data Visualization for Quality Control in NONMEM Data set

Linghui Zhang, Merck Co., Upper Gwynedd, Pennsylvania

ABSTRACT

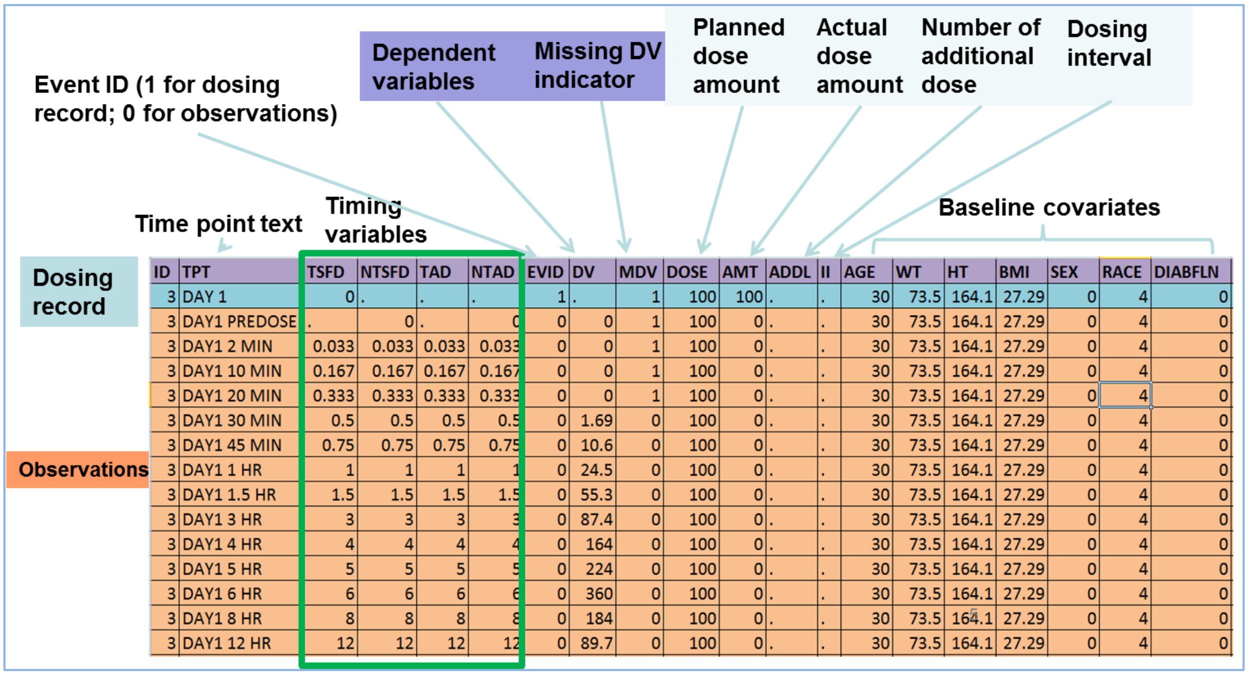
Non Linear **M**ixed **E**ffects **M**odel (NONMEM) data set is widely used for pharmacokinetics (PK) / pharmacodynamics (PD) modeling and simulation, which studies the drug concentration in the body over time (measured in terms of absorption, distribution, metabolism, and excretion [ADME]) and the body's pharmacological response to a drug (measured in terms of adverse events [AE] and efficacies). In a very specific pre-defined format, the NONMEM data set includes a chronological mixture of dosing records, PK/PD observations and covariates of the dosing and observation records. To create NONMEM data sets, it takes tremendous programming efforts for programmers to derive dosing history, order PK/PD observations and merge various types of covariates. The variables required for NONMEM data are often complicated and come from different source data sets. It is a tough challenge to perform data validation and cleaning. Good quality NONMEM data is critical in PK/PD analysis and errors from a small portion of the data can redirect the conclusion of a study. To guarantee the accurate and meaningful PK/PD analysis, data cleaning is essential and crucial for quality control in NONMEM data set production. Graphs are visual summaries of data and very effective to describe essential features than tables of numbers. This paper illustrates some commonly used graphs to virtualize the data errors and questionable records in both raw clinical data and NONMEM data set. Scientific programmers and pharmacometricians with minimal programming skills can apply these graphs to check data issues and examine data thoroughly.

INTRODUCTION

In the field of population pharmacokinetics / pharmacodynamics (pop PK/PD) modeling and simulation, NONMEM software is the leading tool widely used to model and predict the effect of drug on the target population of patients (Shen et al., 2007; Ette et al., 2013). Pharmacokinetics (PK) describes what the body does to a drug by observing the drug concentration in the body over time (). Depending on the drug distribution, PK measurements include the drug concentration in the plasma of blood, urine, target issues and drug concentration in plasma is the most common PK measurement. Pharmacodynamics (PD) is the study of what the drug does to the body. PD measurements might be AEs, biomarkers and efficacies. PK/PD analysis describes the relationship of drug-effects over time, and can be affected by many factors, such as gender, race, food status, biomarkers, etc. These factors affecting PK/PD modeling are called covariates.

To perform pop PK/PD analysis, a NONMEM-ready data set (hereafter also referred to as NONMEM data set) needs to be created based on the modeling specifications (Boeckmann et al., 2011). The data structure of NONMEM data set is very complicated and different from other clinical data sets such as case report tabulation (CRT), Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) data sets. In NONMEM data set, PK/PD observations and dosing events are the key components ordered chronologically and decide the total number of records. The observation records contain the results from the laboratories and the actual dates and times of all the samples taken for PK/PD analysis. Dosing records consist of the actual and nominal dates and times of each drug administration. When a NONMEM data set is generated, the potential covariates influencing on the modeling result will be also included as additional variables in the final data set. Covariates are typically classified as time-independent covariates and time-dependent covariates. The time-independent covariates usually contain basic demographic information, vital sign measurements and lab data collected at the screening phase, etc. They are subject level variables presented as one record per subject in clinical data sets. The time-dependent covariates are recorded over time and usually collected at the time when observation and/or dosing events are taken. A part of typical NONMEM data set is shown Figure 1. In this example, the records consist of a single oral dose followed by a couple of PK measurements. The dose administration is highlighted in blue as the first record. The PK observations are shown in orange after dosing record. The conserved variables in this example are ID, EVID, TSFD, DV, MDV and AMT. The timing variables, TSFD, NTSFD, TAD and NTAD are used to order the records chronically within each ID. Demographic information is also added as covariates.

NONMEM data set creation is usually complicated and time consuming (Arthur et al., 2010). First, the volume of NONMEM data set can be huge due to the number of records and variables. Because population studies are usually performed on large number of subjects, a single NONMEM data set might consist of multiple clinical trials ranging from studies first on human (phase 1) to long-term safety and side-effect (phase 4) of clinical drug development. Based on therapeutic area and study design, variety of study-specific covariables can be included in a single NONMEM data set. Second, the input data usually come in different formats depending on the maturity of data and the development of clinical trial. The source data can come from but not limit in raw data set from clinical data management team, raw excel sheet from lab scientist, CRT, SDTM and ADaM data sets. Therefore, tremendous programming efforts are needed to pool data from multiple clinical studies, merge data coming in diverse formats, combine and validate records in a single SAS output data set. Third, NONMEM data set production is usually driven by the pharmacometricians based on the modeling analysis plan (MAP), but ad hoc requests come in many scenarios, i.e., interim analysis for efficacy concerns, and data querying by Medical Monitoring Committees for safety reasons. In addition, pharmacometricians may update NONMEM data specification frequently due to modeling purpose. Data refreshes for the purpose of data set development also make reproduction, debugging, and revalidation time consuming. All of these causes the high complexity and time-consuming in the generation of NONMEM data set.


Event ID (1 for dosing record; 0 for observations)
Dependent variables
Missing DV indicator
Planned dose amount
Actual dose amount
Number of additional dose
Dosing interval
Time point text
Timing variables
Baseline covariates
Dosing record
Observations

ID	TPT	TSFD	NTSFD	TAD	NTAD	EVID	DV	MDV	DOSE	AMT	ADDL	II	AGE	WT	HT	BMI	SEX	RACE	DIABFLN
3	DAY 1	0	.	.	.	1	.	1	100	100	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 PREDOSE	.	0	.	0	0	0	1	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 2 MIN	0.033	0.033	0.033	0.033	0	0	1	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 10 MIN	0.167	0.167	0.167	0.167	0	0	1	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 20 MIN	0.333	0.333	0.333	0.333	0	0	1	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 30 MIN	0.5	0.5	0.5	0.5	0	1.69	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 45 MIN	0.75	0.75	0.75	0.75	0	10.6	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 1 HR	1	1	1	1	0	24.5	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 1.5 HR	1.5	1.5	1.5	1.5	0	55.3	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 3 HR	3	3	3	3	0	87.4	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 4 HR	4	4	4	4	0	164	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 5 HR	5	5	5	5	0	224	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 6 HR	6	6	6	6	0	360	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 8 HR	8	8	8	8	0	184	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0
3	DAY1 12 HR	12	12	12	12	0	89.7	0	100	0	.	.	30	73.5	164.1	27.29	0	4	0

Figure 1. Example of records and variables in a NONMEM data set.

Data cleaning is the essential key to guarantee the good quality of NONMEM data set and especially required when integrating heterogeneous data sources (Bonate et al., 2012). Many data issues can be found in the source data and the final NONMEM data set, such as the incorrect observation results, incomplete data collections, improperly format, missing information, inconsistencies, etc. Even though data management team, programmers and pharmacometricians examine data for flaws by using rules, algorithms, and look-up tables systematically, some hidden data issues are still present in the final production and impact the accurate and meaningful analysis of PK/PD modeling and simulation.

Graphical methods provide visual representations of data and are more quickly and completely to show the general trend and relationships of variables. In pop PK/PD analysis, graphs play an important role in showing the drug's concentration profiles and drug-effect relationships. Moreover, it also provides insight for the analyst into the data issues and errors. This paper applied some common used graphs to detect the data issues of PK/PD observations and covariates. Scientific programmers and pharmacometricians with minimal programming skills can also apply SAS/GRAPH® to visualize data issues in the early stage of clinical trials.

CASE 1: EXAMINE PK CONCENTRATION BY INDIVIDUAL PK-TIME PLOT

The individual PK-Time plot provides a quick view of the general trend of the drug concentration over time. It is very effective to check the potential data errors of PK concentration at the early stage of the study by examining individual PK-time plot. The sample codes below can simply plot the dose and PK observation over time and the output graphs are shown in Figure 2.

```
PROC SGPANEL DATA = nm01 ;
       WHERE evid = 1 ; * Plot the dosing records ;
       PANELBY an / ONEPANEL COLUMNS = 2; * show in onepanel with two columns ;
       SCATTER x = tsfd y = amt / GROUP = peri ; * group by period ;
       SERIES  x = tsfd y = amt / GROUP = peri ;
RUN ;
PROC SGPANEL DATA = nm01 ;
```

```
        WHERE evid = 0 ; * Plot PK concentrations ;
        PANELBY an / ONEPANEL COLUMNS = 2;
        SCATTER x = tad y = dv / GROUP = peri ;
        SERIES  x = tad y = dv / GROUP = peri ;
RUN ;
```

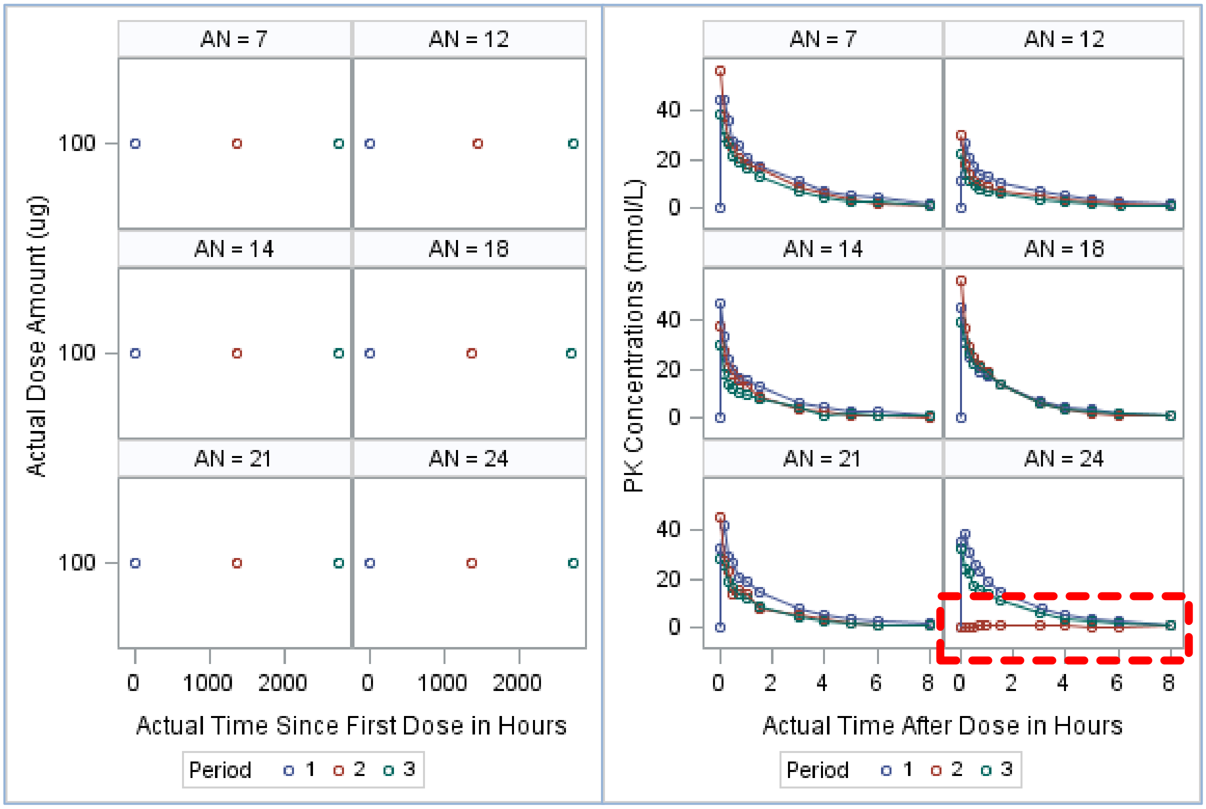


Figure 2. Dose vs Time and PK vs Time show Dose-PK relationship.

As shown in the plot of Actual Dose Amount vs the Actual Time Since First Dose (TSFDA) in the left panel of Figure 2, each subject in this treatment group was administrated with a single IV dose in each of the three study periods. Therefore, the PK concentration over the Actual Time After Dose (TAD) will be expected as a nice curve with a long declined tail of distribution and elimination phase. Any spots away from the curve will be recognized as questionable records. For example, the PK concentrations are close to zero and PK curve is flat as a horizontal line in the period 2 of subject AN24 (highlighted by the box of red dotted line). The distribution and elimination phase were not observed in this period. These unusual PK concentrations should be examined further and reported to pharmacometrician.

Besides plotting the PK concentration over TAD, we can also plot PK concentration over the TSFDA to examine the PK profile thoroughly:

```
PROC SGPANEL DATA = nm02 NOAUTOLEGEND ;
        WHERE evid = 0 & reqm = 2 ;
        PANELBY an / ONEPANEL ROWS = 3 ;
        SCATTER x = tsfd y = dv ;
        SERIES  x = tsfd y = dv ;
RUN ;
```

In this study, six subjects were administrated with four oral doses based on protocol and PK observations were collected before and after each dose. As expected, there're four absorption peaks in the PK profile of each subject in Figure 3 (E.g., AN4 and AN16). However, some subjects have either extra or missing absorption peaks, which are not present in a typical PK curve as expected. For example, subject AN1 has no absorption peak after the second dose; there's an unexpected PK concentration peak for subjects AN9 and AN11 between 3000 and 3500hr of TSFDA since there's no additional dose at the mean time upon protocol. These PK concentrations in subject AN1, AN9 and AN11 should be reported to pharmacometrician. In addition, the missing records in PK curve indicated the early discontinuation in subject AN15. Therefore, the termination status needs to be confirmed by data management team. Otherwise, these missing collections should be confirmed by lab specialist.

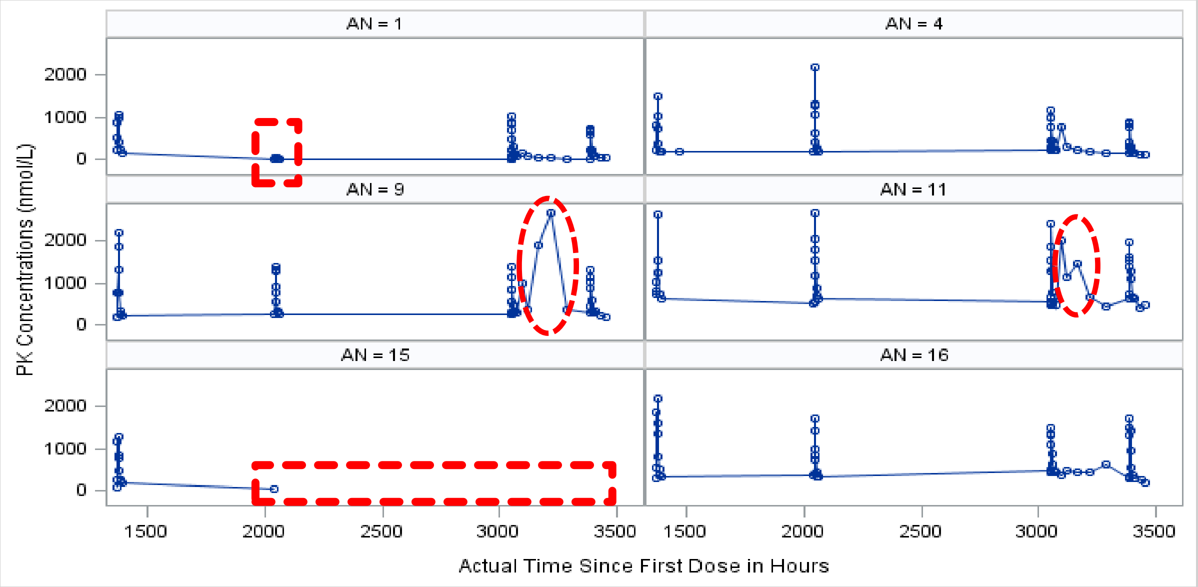


Figure 3. PK Profile Over Actual Time Since First Dose.

CASE 2: EXAMINE PK CONCENTRATION BY POPULATION PK PLOT

It is time consuming to detect each individual PK-Time plot when sample size is large, such as the individual PK profiles from phase 2b and phase 3 studies. Alternatively, we can examine the overall distribution of PK concentration and check the outliers in population level. It is efficient to oversight the range of PK concentrations and statistical outliers over this range are suspected as the questionable records. Three useful graphs, histogram, box plot and scatter plot, will be applied to examine the overall distribution of PK concentration.

In a phase 3 study with over 800 patients in treatment group, subjects took one tablet of the study medicine daily with meal and the trough plasma PK samples were collected after 12-hour fast and voiding in eight visits during the treatment phase. Due to the unusually long half-life of the drug, the follow-up PK concentrations were also collected every three months up to two years after treatment termination. Obviously, it's not feasible to go over individual PK profile one by one for all of the 800 subjects. Because the trough PK is the lowest concentration that a drug reaches before the next dose is administered, we should expect the lower PK concentrations in the visits of treatment phase. Then, any outliers of unusual high concentration might be questionable PK concentrations. The graphical methods showing the range of population distribution, such as histogram, box plot, scatter plot can be the good tools.

The sample code below was used to generate the histograms of trough PK in treatment phase:

```
PROC SGPANEL DATA = nm02 ;
       WHERE visit <= 26 ; * visits in treatment phase ;
       PANELBY visit / ONEPANEL ROWS = 2 ;
       HISTOGRAM dv ;
RUN ;
```

As we can see from the histogram of PK distribution per visit in Figure 4, the trough PK concentrations under the treatment phase range from 0 to 5000 nmol/L with a few outliers over 5000 nmol/L. The high concentration doesn't merely indicate the questionable PK records because PK concentrations vary cross subjects and can be affected by many covariates. To rule out high concentration caused by the variety response of individual effect and covariates, the PK records of subjects having PK concentration over 5000 nmol/L were pooled and the box plots (Figure 5) of PK concentration for each subject are used to examine the outliers over 1.5 IQR (interquartile range) within each subject. Four subjects had PK concentration greater than 1.5 IQR under treatment phase. The visit number is used to label the outliers of each box. Further inspections are necessary for these four subjects. The sample codes below are used to pool the questionable records and create box plot by subject.

```
PROC SQL ;
       CREATE TABLE hightrconc AS
       SELECT * FROM nm02 WHERE an in
              (SELECT distinct an FROM nm02 WHERE dv > 5000) ;
QUIT ;
PROC SGPANEL DATA = hightrconc ;
```

```
        PANELBY an / ONEPANEL ROWS = 1 ;
        VBOX dv / DATALABEL = visit ; * label the outliers by visit ;
RUN ;
```

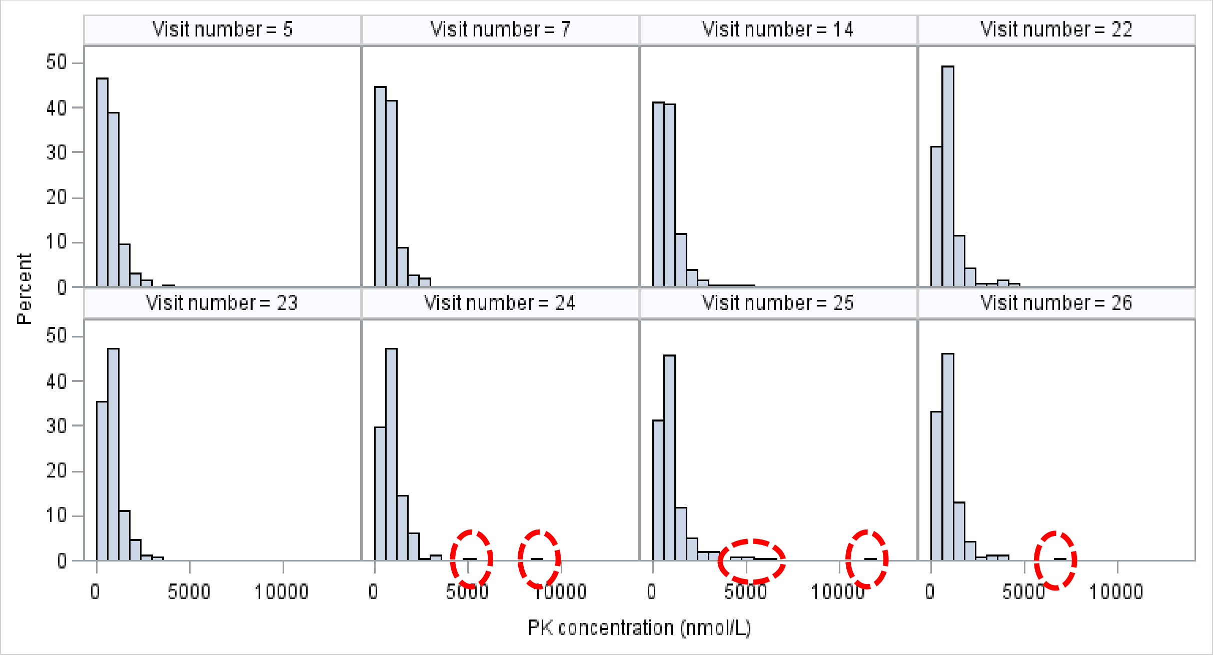


Figure 4. Distribution of PK concentration by visit in treatment phase.

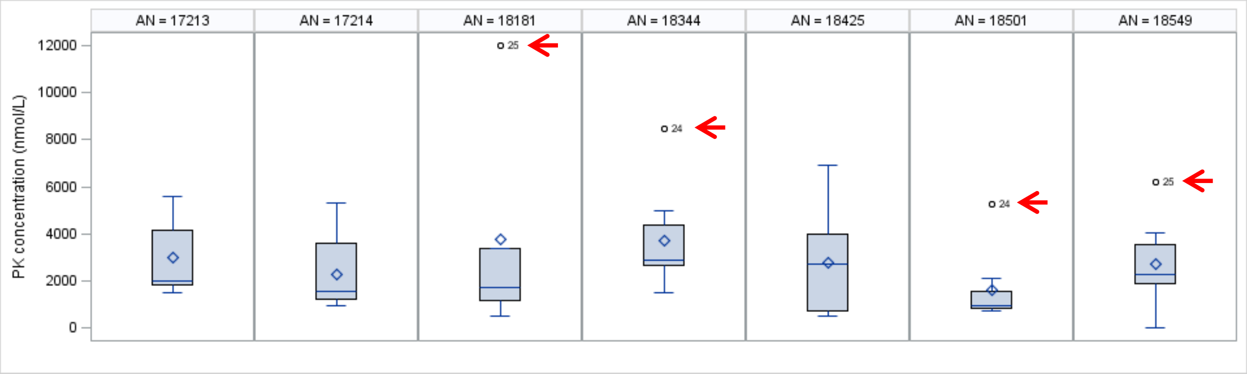


Figure 5. Box plot of PK concentration from the subjects with concentration higher than 5000 nmol/L.

In some cases, it's very common in pop PK/PD analysis to use the natural logarithm of pk concentration to model the PK profile and PK-PD relationship. However, the logarithm of PK concentration is not a good source data if the primary purpose is to catch concentration outliers. In statistics, the transformation of logarithm is widely applied to normalize and centralize the uneven skew distribution. Thus, the unusual concentrations or outliers might not be noticed after being converted to logarithm. In addition, only the positive values have logarithm, if the concentration is zero, it will be missing if logarithm is used to generate PK profile. Therefore, raw data without any mathematical conversion is favored to create graphs for the overview of data distribution.

The unusual PK concentrations collected in the follow-up phase can be also detected by graphical methods. In this study, plasma PK was tested every three months up to two years after treatment termination. Due to the long elimination half-life of the drug, the PK curve enters a slower phase of disappearance after treatment termination. Therefore, the slower decline is present in the PK profile due to the irreversible drug elimination or clearance. Because PK curve enters decline phase, histogram by visit is not a good idea to identify the unusual high concentration in elimination. Alternatively, we can create scatter plot of PK concentrations collected in the two consecutive visits, such as the PK concentration at a certain visit over the concentration at the visit prior to the certain visit.

```
* lag PK concentration by tsfd within each subject ;
```

```
PROC SORT DATA = nm02 ;
       BY an tsfd ;
RUN ;
DATA nm02 ;
       SET nm02 ;
       BY an tsfd ;
       LABEL lst dv = "PK conc. at the previous visit"
                  dv = "PK conc. at the current visit" ;
       lst dv = lag(dv) ;
       if first.an then lst_dv = 0 ;
RUN ;
* create scatter plot of current pk conc. vs last pk conc.
PROC SGPANEL DATA = nm02 noautolegend ;
       WHERE visit > 26 & lst dv < dv & lst dv > 0 ; * follow-up phase ;
       PANELBY visit / onepanel columns = 4 ;
       SCATTER x = lst dv y = dv ;
       LINEPARM x =0 y = 0 slope = 1 / LINEATTRS = (color = blue) ;
       LINEPARM x =0 y = 0 slope = 1.5 / LINEATTRS = (color = orange) ;
       LINEPARM x =0 y = 0 slope = 2 / LINEATTRS = (color = red) ;
RUN ;
```

The reference line of equal concentration between two consecutive visits is added to discriminate the PK concentrations higher than that collected at previous visit. In the scatter plots shown in Figure 6, the PK concentration at the current visit vs the visit prior to the current one (current visit vs last visit) is created and three reference lines are added to indicate 1 (blue), 1.5 (orange) and 2 (red) times of current concentration over the last one. For example, if the current concentration is equal to or lower than the last test, the spot locates between the 45-degree line (blue) and horizontal axis. Otherwise, the spots presenting between the 45 degree line and vertical axis suggest the increase of PK concentration after treatment termination, which is unexpected in the elimination phase. In this plot, only the spots between the vertical axis and the 45-degree line are displayed since all the spots on the other side of 45 degree line show the decrease in PK concentration after last visit. Similarly, the spots over the orange and red reference lines can be recognized as the high panic concentration because the values are 1.5 and 2 times of last test.

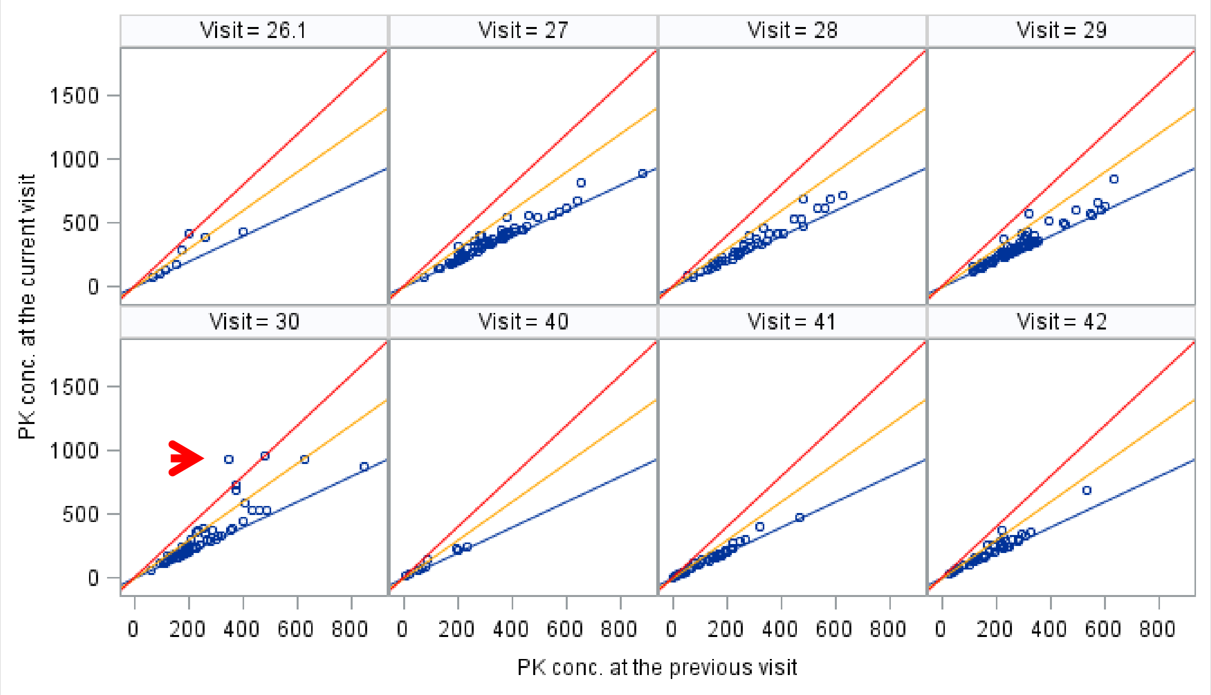
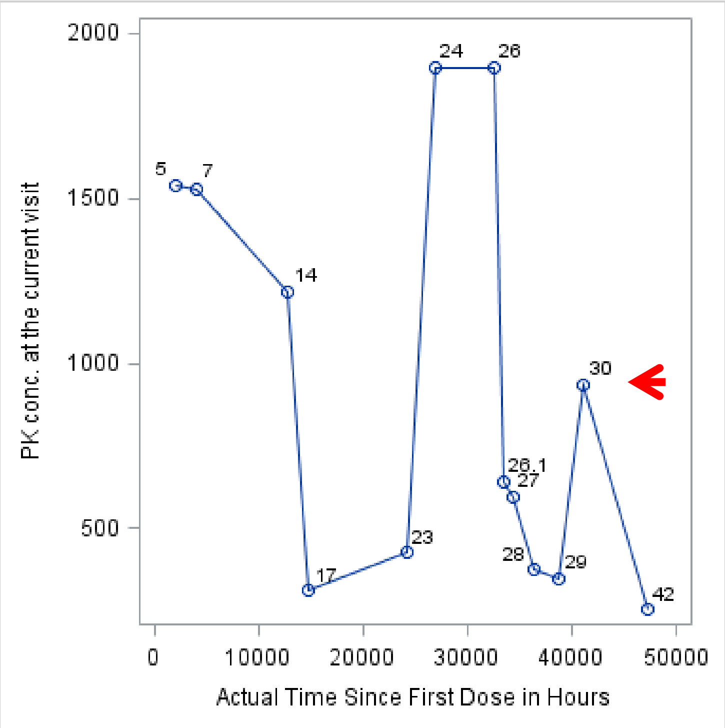
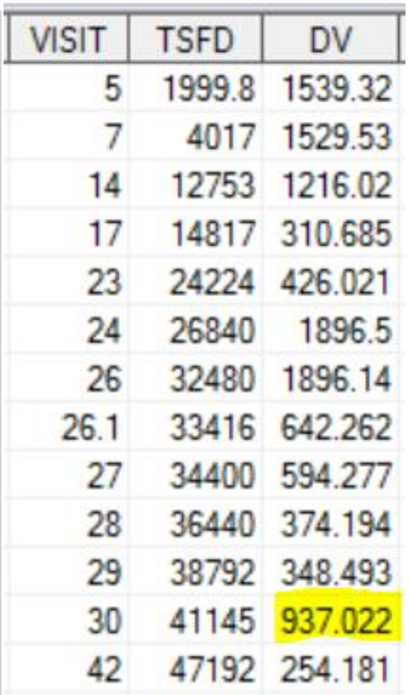


Figure 6. Scatter plot of PK concentrations in two consecutive visits.

For instance, there’s one spot over the red line at visit 30 indicating that the PK concentration tested at visit 30 is over two times of PK concentration tested at visit 29 for the certain subject. All the observations of subjects with PK concentration located between the vertical axis and red line were pooled to data set dv2. Then the PK curve of subject 17695 in dv2 is shown in Figure 7 as an example. We can see the concentration bump at visit 30, which is found to be the questionable record.

```
PROC SQL ;
       CREATE TABLE dv2 AS
       SELECT * FROM nm02 WHERE an in
              (SELECT distinct an FROM nm02 WHERE dv > lst_dv*2 & lst_dv > 0 & visit >
              26 )
       ;
QUIT ;
PROC SGPLOT DATA = dv2 NOAUTOLEGEND ;
       WHERE an = 17695 ;
       SCATTER x = tsfd y = dv / DATALABEL = visit ;
       SERIES x = tsfd y = dv ;
RUN ;
```





VISIT	TSFD	DV
5	1999.8	1539.32
7	4017	1529.53
14	12753	1216.02
17	14817	310.685
23	24224	426.021
24	26840	1896.5
26	32480	1896.14
26.1	33416	642.262
27	34400	594.277
28	36440	374.194
29	38792	348.493
30	41145	937.022
42	47192	254.181

Figure 7. Example of individual PK profile to show the concentration bump after treatment termination

CASE 3: CHECK THE CONSISTENCY OF ACTUAL TIME AND NOMINAL TIME

In NONMEM data set, the observations and dosing administrations are ordered based on the chronological sequence of these events within each subject. Thus, timing variables are critical for accurate and meaningful analysis of PK/PD modeling and simulation. Two types of timing variables are collected, nominal time and actual time. A nominal time is the planned sampling or dosing time. For instance, the planned time for PK/PD observation can be “PREDOSE”, “12 hours POST DOSE”, the scheduled time for dose administration can be “Period 2 Day 3”. In addition to the nominal time, the actual date and time is also recorded by the healthcare professional when an event is conducted. Thus, the actual date and time can be used to calculate the actual time after dose for a special event. It is important to examine the consistency between the actual and nominal time in case sample and dose dates and times are messed up.

Two graphical methods are applied to evaluate the consistency between nominal and actual time. One method is to check the distribution of the difference between nominal time and actual time.

```
DATA nm03 ;
       SET nm03 ;
       d = abs(tsfda-tsfdn) ;
RUN ;
PROC SGPANEL DATA = nm03 ;
       PANELBY npanel peri / ONEPANEL UNISCALE = column ROWS = 1 ;
```

```
        HISTOGRAM d / SCALE = count ;
RUN ;
```

In Figure 8, the absolute difference betw een actual and nominal time since first dose are calculated and the histogram of difference by visit is plotted. We can tell from the distribution that, the differences betw een actual and nominal time are consistent in the most of the cases and are closed to zero. How ever, the huge differences at panel 2 and 3 are observed, w hich indicates the inconsistency betw een nominal and actual time.

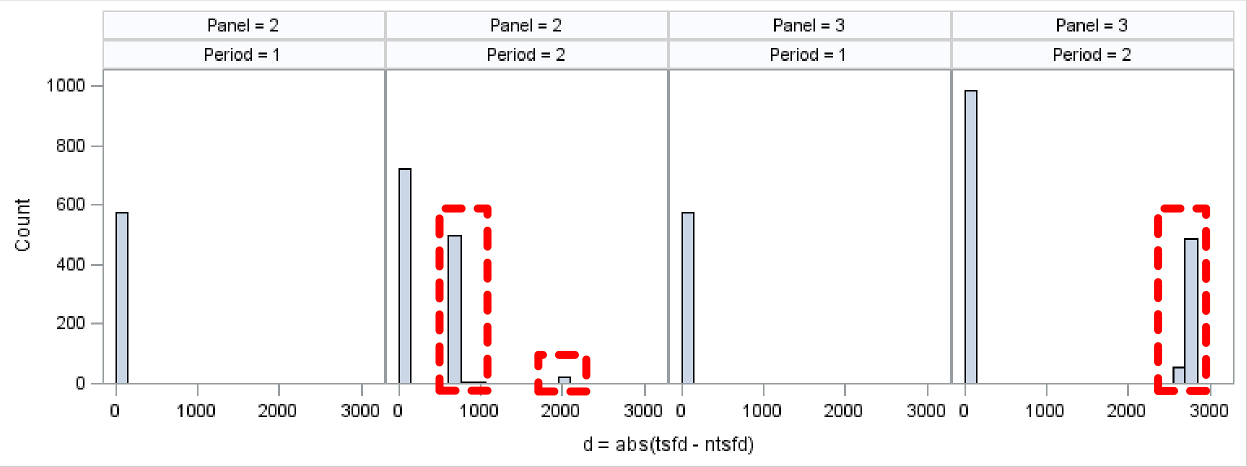


Figure 8. Distribution of absolute difference between actual time and nominal time.

The other method is to create the scatter plot of the tw o types of times directly and look for the spots aw ay from the 45 degree line w hich indicates the actual time is identical to the nominal time.

```
PROC SGPLOT DATA = nm03 ;
        WHERE npanel = 2 & peri = 2  ;
        SCATTER x = ntsfd y = tsfd / GROUP = regm ;
        LINEPARM x=0 y=0 slope=1 ;
        XAXIS GRID ;
        YAXIS GRID ;
RUN ;
```

As Figure 9 show n, most of the spots are located along the 45-degree diagonal line, w hile a few spots are above or under this reference line. The spot presenting at 10000 hours of nominal time needs to be further inspected.

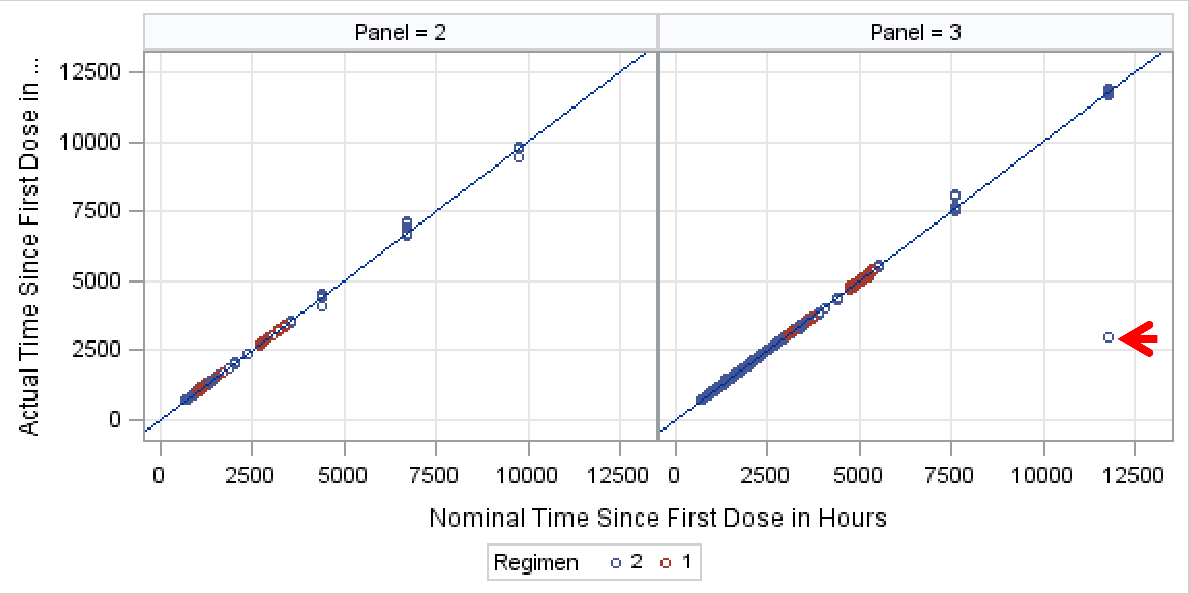


Figure 9. Scatter plot of actual time vs nominal time to show the discrepancy between the two types of time.

CONCLUSION

With powerful data visualization methods, programmers can catch the data issues in the early stage of clinical trial. SAS graphical methods also provide timely visual access for data exploration to pharmacometricians. Changing data types and repeatedly transferring data between functional teams are not required since SAS is the major data type used in clinical trial data sets. User-friendly SAS graphical tools require minimal graphic programming skills from SAS programmers and pharmacometricians.

REFERENCES

Shen, D., Lu, Z., 2007. Population Pharmacokinetics Studies with Nonlinear Mixed Effects Modeling.

Ette, E.I., Williams, P.J., 2013. Pharmacometrics: The Science of Quantitative Pharmacology.

Boeckmann, A.J., Sheiner, L.B., Beal, S.L., 2011. NONMEM Users Guide.

Arthur Collins, Mark Peterson, Greg Silva. 2010. Streamlining the PK/PD data transfer process. Pharmaceutical Programming, June 2010, 24-28.

Bonate, P.L., Strougo, A., Desai, A., Roy, M., Yassen, A., Walt, J.S., Kaibara, A., Tannenbaum, S., 2012. Guidelines for the Quality Control of Population Pharmacokinetic–Pharmacodynamic Analyses: an Industry Perspective. AAPS Journal, v.14(4); 2012 Dec, PMC3475847.

CONTACT INFORMATION

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

Name: Linghui Zhang
Enterprise: Merck Co.
Address: 351 N Summeytown Pike, North Wales, PA 19454
Work Phone: (267)305-6747
Fax: (267)305-6538
E-mail: linghui.zhang@merck.com