Efficient quality review for modeling input dataset

Vincent Buchheit, Pui Tang, Aurélie Gautier, Thomas Dumortier, Grégory Pinault

Modeling & Simulation, Novartis Pharma AG, Basel, Switzerland

Purpose

The poster's purpose is to share our experiences and present our views on quality review for modeling input dataset (MID). We wish to encourage and enhance feedback and collaboration from colleagues.

Introduction

Drug development is a succession of clinical trials, where statisticians and programmers play a major role in data reporting and data analysis. In addition, pooling data across studies, within a compound, or across compounds is becoming nowadays a routine activity in most pharmaceutical companies, also for the Modeling and Simulation (M&S) Programming Group at Novartis. Tools and methodologies are developed[1] to facilitate data pooling, we are still facing the following challenge: "How can we perform an efficient quality review of our pooled modeling input dataset, and therefore *validate* the data?"

Historically, the *double* programming process is routinely applied in pharmaceutical industries. An independent programmer re-produces the same dataset based on data specifications. The validation is completed when both datasets match. It takes up to several weeks before completion. This Quality Control (QC) method is adequate to ensure that the data-generating program does what it is supposed to do. However, it does not guarantee that the data is scientifically accurate.

Modeling input dataset composition

We define the MID as an array containing all the data for a modeling activity of any kind. A modeling file (**Figure 1**) is composed of various variables which can be classified into 5 types[1]:

- Identification variables - e.g. ID variable...
- Time variables - e.g. Time since the very first dose, Time since the very first event, Elapse time...
- Events [dose administrations, pharmacokinetic (PK), pharmacodynamic (PD), biomarkers, clinical scores...] - e.g. CMT, DV variables...
- Time independent covariates - e.g. Age, Sex, Body Surface Area (BSA) at baseline...
- Time dependent covariates (which vary with time) - e.g. Serum creatinine, Creatinine clearance...

Figure 1. *Example of a modeling file*

id	time	tad	cmt	lidv	mdv	evid	amt	dose	age0	bmi0	wt0	sex	race
1	-25.55	-25.55	3	0.999228	0	0	0	0	64	24	77	1	1
1	-25.55	-25.55	4	0.314407	0	0	0	0	64	24	77	1	1
1	-0.033	-0.033	2	0	0	0	0	0	64	24	77	1	1
1	0	0	1	0	1	1	100	100	64	24	77	1	1
1	0.5	0.5	2	0.446197	0	0	0	100	64	24	77	1	1
1	1	1	2	0.384585	0	0	0	100	64	24	77	1	1
1	1.5	1.5	2	0.222677	0	0	0	100	64	24	77	1	1
1	2	2	2	0.124808	0	0	0	100	64	24	77	1	1
1	3	3	2	0.954878	0	0	0	100	64	24	77	1	1
1	4.017	4.017	2	0.067284	0	0	0	100	64	24	77	1	1
1	6.017	6.017	2	0.007714	0	0	0	100	64	24	77	1	1
1	8	8	2	0.975366	0	0	0	100	64	24	77	1	1
1	24.033	24.033	2	0.283091	0	0	0	100	64	24	77	1	1
1	24.05	0	1	0	1	1	100	100	64	24	77	1	1
1	46.45	0	1	0	1	1	100	100	64	24	77	1	1
1	70.45	0	1	0	1	1	100	100	64	24	77	1	1
1	94.45	0	1	0	1	1	100	100	64	24	77	1	1
1	118.45	0	1	0	1	1	100	100	64	24	77	1	1
1	142.45	0	1	0	1	1	100	100	64	24	77	1	1
1	142.45	0	3	0.310678	0	0	0	100	64	24	77	1	1
1	142.45	0	4	0.972884	0	0	0	100	64	24	77	1	1
1	167.333	24.883	2	0.908268	0	0	0	100	64	24	77	1	1

To create a MID you often have to combine data from different sources[2] (dose history, PK, demographics, randomization, biomarkers...) and sometimes across studies and across compounds. You can end up having one single file that contains millions of observations and several columns (10-100). Usually the data cleaning is done by data managers during each trial. They verify data by type of exam (verify dose history alone, verify demographics alone...), but they do not perform any cross validation within a subject (check the consistency of pharmacokinetic data according to the complete dose history)... Therefore it is quite frequent to identify inconsistencies at that level in a MID.

First level of verification

Before performing any cross validation, it is essential to look at each key variable alone (MID columns). There are multiple sources of error:

- Unit problem – dose, PK, PD, time independent covariates, time-dependent covariates
  - not correctly recorded in the database: values expressed in mg/dL but unit is g/dL
  - conversion factor used to convert all raw data to the same unit across all centers and studies could sometimes be wrong: convert mg/dL to umol/L using 8.4 instead of 88.4
  - different units used across all studies but not identified during the MID creation
- Outliers – identification variables, time variables, dose, PK, PD, time-independent covariates, time-dependent covariates:
  - data entry error: 13006 instead of 130.06
  - variables or dataset used not appropriate, including redundant but not consistent information (time of dose recorded in different datasets but values are different)
  - algorithm to replace missing values inadequate
  - programming error

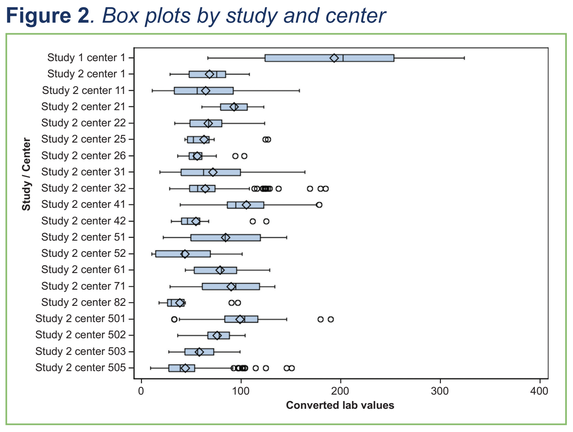
More specifically for the identification variables (patient ID) and time independent covariates (gender, ethnicity, age at baseline....), we want to ensure that values are unique for each patient. We need to make sure that values do not change over time (for example unique body weight by patient). Attention must be made if some of the patients were enrolled in different studies. While there is no easy way to identify such data issues in one go, a summary table (**Table 1**) (including number of observations, number of missing observations, minimum, 25% percentile, median, 75% percentile, maximum) and a series of boxplots will provide an accurate description of your data.

Table 1. *Summary table by study*

obs	STUDY	n	nmiss	min	max	q1	median	q3
1	study 1	132	0	66.6000	324.000	123.800	202.000	253.050
2	study 2	2070	9	8.9075	189.940	41.918	60.060	88.813

If it is a pooled analysis, we suggest to produce the summary table and the boxplots by study, or by study and center (**Figure 2**); if it is only a single study, by centers only. You can then identify potential issues and try to understand the reasons for them. In the example above, it is obvious by just looking at the summary table that the laboratory values for study 1 are much higher compared to those in study 2. After ensuring that the problem was not coming from the program that generates the MID, we realized that the unit associated with the raw data was not correct, for the entire study.

Figure 2. *Box plots by study and center*



As soon as you identify data issues, you have to understand where the issue could come from. Is it coming from the program that generated the modeling dataset? Is it coming from the raw data? Is it coming from the data specifications?

Assuming your MID includes continuous variables (for examples laboratory data), we always convert them to the same unit (often SI unit). While it is important to look at the converted values, it is also important to look at the variable distribution before the conversion. If just 10% of your values are suspicious, it may be difficult to observe it in the converted values. In addition, having summary statistics by units could also be used to estimate and replace missing units.

For the time-dependent covariates (body weight by visit, serum creatinine over time...), one way to do it is to plot each time-dependent covariates for each patient (**Figure 3**). While it is easy to do, it could be time consuming to review them all. An alternative is to produce series of boxplots by defined scheduled timepoint (**Figure 4**).

Figure 3. *Time-dependent covariate over time*

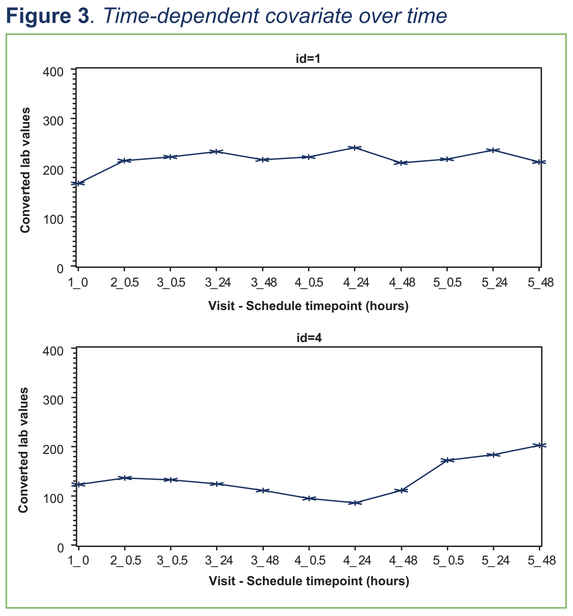
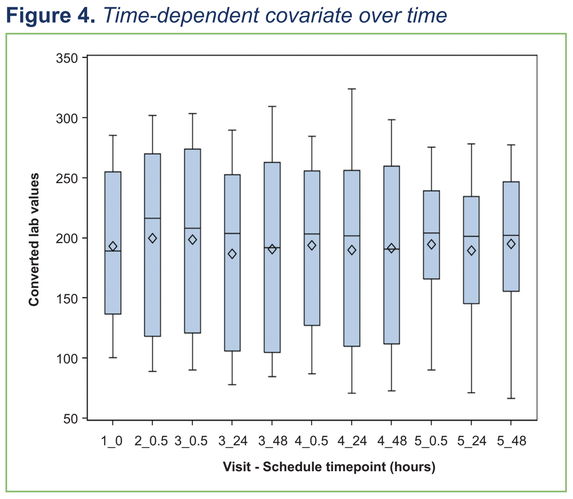


Figure 4. *Time-dependent covariate over time*



It is also important to look at the dose level and dose regimen observed in the MID and compare it to the treatment description in the protocol.

Second level of verification

Once the first level of verification is completed, it is time to look at some combined data. In the previous section we explore the dose history alone, the PK data alone but we haven't looked at the PK combined with the dose history for example. While we concentrate only here on the PK data, the same principle apply to the PD data as well.

There are several ways to explore the PK and dose history data together. You can plot all PK data and all dose history in one plot (**Figure 5**), you can plot all PK data by elapsed time by dose level (**Figure 6**). Assuming your PK is linear, you can also plot PK/dose by elapsed time. The latter allows you to look at all PK data in one plot.

Figure 5. *PK dose history and dose administration over time*

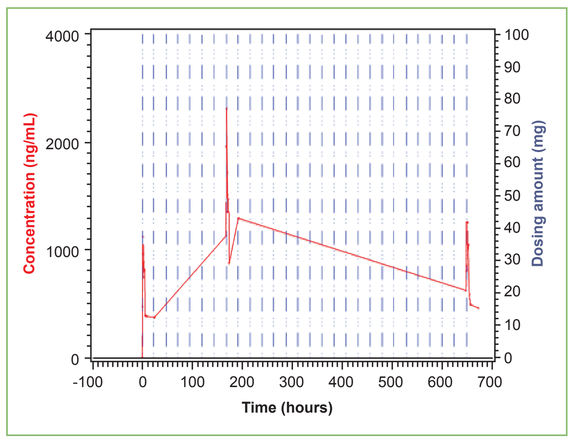
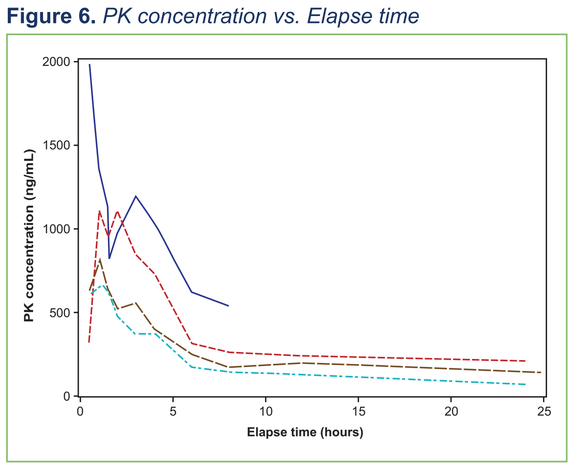


Figure 6. *PK concentration vs. Elapse time*



Data collection is defined in the protocol by visit and timepoint. Comparing the scheduled timepoint (stated in the protocol) with the actual data is often a meaningful source of information. It allows you to detect deviation from protocol, data entry error and it also allows confirmation of some of the imputations. For this purpose you can plot the elapsed time by the protocol timepoint (**Figure 7**).

Figure 7. *Scheduled timepoint vs. Elapse time*

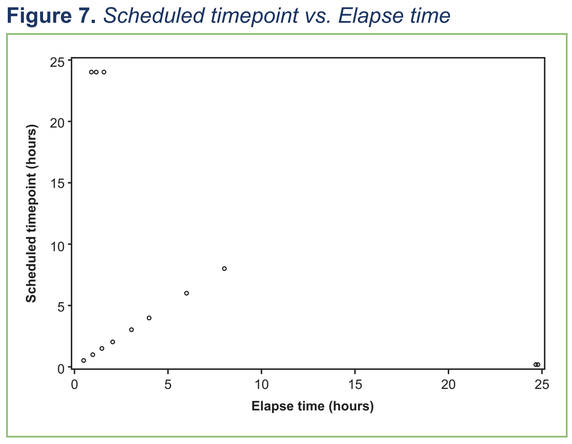
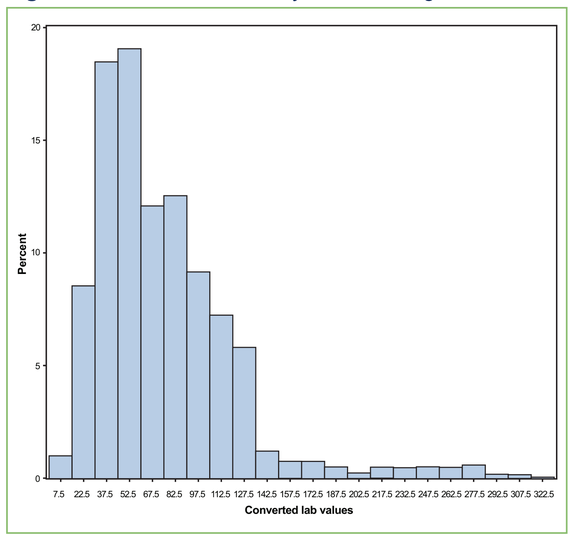
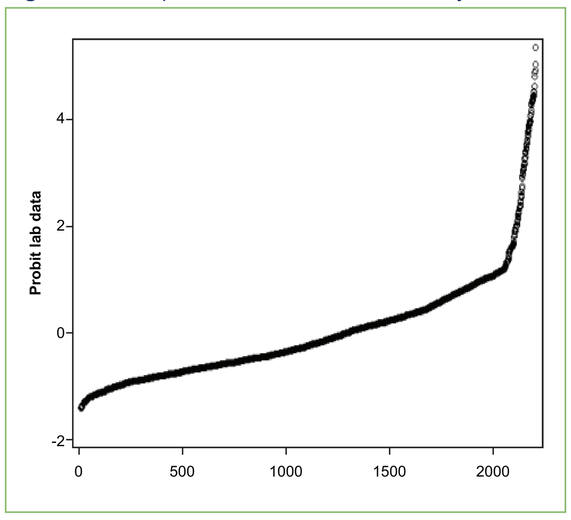


Figure 8. *Converted laboratory values histogram*



Finally, you may want to explore the possibility of a multimodal distribution. The simplest method is a histogram. However the histogram (**Figure 8**) can be misleading, depending on the choice of cell division[3]. A more robust method is the probit plot (**Figure 9**). The probit function is the inverse of the standard normal cumulative distribution function.

Figure 9. *Probit plot for the converted laboratory values*



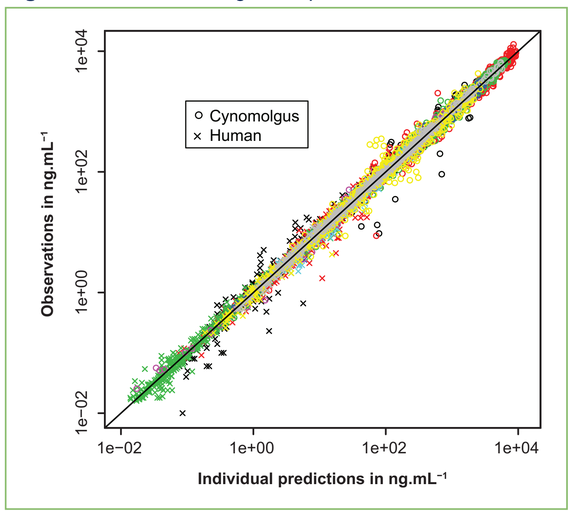
Re-use of a qualified model

This approach is based on two main assumptions:

- The model has already been qualified
- Data is comparable to data previously used for the qualification of the model (e.g.: a new study for a given compound, a new compound with a similar mechanism of action and kinetics...)

Based on the previous model, either a post-hoc estimate (e.g.: MAXEVAL=0 with NONMEM) can be performed, or the data can be fitted on the qualified model. In any case an evaluation of estimates or an evaluation of the goodness of fit has to be performed. Below, a model has been fitted on 8 new different compounds (having linear PK, with similar mechanism of action) in 2 different species (cynomolgus and human). A standard diagnostic plot showing dependent variables observed versus individual predictions with different color per compound (**Figure 10**) "validate" the data quality.

Figure 10. *Standard diagnostic plot*



This approach is a good complement to what has been described earlier. The advantages of this approach are quite interesting. It is very efficient since inconsistencies of data are easily spotted; and time saving because when a model takes a long time to run, a post-hoc estimate shows inconsistencies from the beginning.

Conclusion

From our experience, the double programming alone does not ensure data quality for model-based analyses. Therefore, it is important to review the data before and after any manipulations by means of graphics and summary tables. This minimizes errors and provides accurate results that can be used for the model building. The quality of the modeling input to clinical team depends on the quality of the data[4], therefore it is important to maximise the quality of the MID.

Acknowledgment

We would like to thank Jean-Louis Steimer and Hugh McDevitt who helped us to create this poster. We would like to thank all our colleagues from M&S at Novartis who provide an exciting and challenging work experience.

This study was supported by Novartis Pharma AG, Basel, Switzerland.

References

1. Pinault, G. et al., A structured approach to industrialize the data sourcing to support model based drug development, PAGE 20 (2011) Abstr 2015 [www.page-meeting.org/?abstract=2015].
2. Fotteler B. Et al., From blood flow to data flow: How to build a data set for NONMEM evaluation. An industrial experience. COST-B1 Conference "The population approach: measuring, and managing variability in response. concentration and dose", Geneva (CH), 12–14 February 1997.
3. Jackson, P. R. et al., Testing for bimodality in frequency distributions of data suggesting polymorphisms of drug metabolism-histograms and probit plots, Br. J. clin. Pharmac. (1989), 28, 647–653.
4. Aarons, L. et al., Role of modelling and simulation in Phase I drug development.

Poster presented at the Population Approach Group Europe (PAGE), June 7–10, 2011, Athens, Greece.