

A structured approach to industrialize the data sourcing to support model based drug development

Gregory Pinault, Georges Ette, Hugh McDevitt, Vincent Buchheit

Modeling & Simulation, Novartis Pharma AG, Basel, Switzerland



Background

The value of the Modeling and Simulation (M&S) group has been recognized within the organization¹. Development of new drugs now requires more systematic and more frequent model based insights from M&S². Input is given at all stages of the drug development lifecycle: from very early exploratory research till post registration activities. Models are built over aggregations of all relevant available data and updated on a regular basis as new data become available. At any point in time during drug development (eg. IND, end of phase II) the resulting data analysis can be used for decision making and / or for submission to Health Authorities. Each step presents different challenges in terms of timelines, data access, etc. Nonetheless every prepared modeling dataset has to comply with the regulatory requirements and good clinical practices including the audit trail and the detailed specifications. Even though the clinical teams include pharmacokinetic (PK) and pharmacodynamic (PD) data in the clinical database, the data environment remains designed for more conventional statistical approaches and triggers the usual challenges³. Indeed, the clinical databases are designed to answer different needs (see Table 1). The data are organized amongst different panels with differences in their format between different databases. So data are then reconciled before they can be used in a non linear mixed effect model (eg. Nonmem). Moreover, most of the clinical timelines are set by the clinical teams in reference to the production of the traditional tables, listings and graphics from the statistics group.

Covariate variables: can be of two types:

- Covariates: what are the relevant parameters to consider with the interpretation of the measurement? Some may be time-dependent and others time-independent. E.g. age at baseline, Body Mass Index, Serum creatinine, etc...
- Flags: Are there any particularities on that individual or event to be aware of and to either keep or reject the record for the analysis? Some may be time-dependent and others time-independent. E.g. flag positive pre very first dose drug plasma concentration, flag missing creatinine level at baseline, etc...

Figure 1. *Proposed classification of variables of modeling data files*

ID variables	TIME variables	EVENT variables	COVARIATE and FLAG variables				
STUDY CENTER SUBJ ID	TIME TIM2 NT	CMT EVID LIDV AMT	MDV AGE HTO COU IGN1 IGN2				
12103 1 5017 1	0.00 0.00 0.00	2 0 0.00 0	1 49.78 187 DEU 1 0				
12103 1 5017 1	0.25 0.00 0.00	1 1 0.00 50	1 49.78 187 DEU U U				
12103 1 5017 1	0.50 0.25 0.25	2 0 65.00 0	0 49.70 107 DEU 0 0				
12103 1 5017 1	0.50 0.25 0.25	2 0 47.00 0	0 49.78 187 DEU 0 0				
12103 1 5017 1	1 25 1 00 1 00	2 0 24.00 0	0 4978 187 DEU 0 0				
12103 1 5017 1	2.25 2.00 2.00	2 0 19.10 0	0 49.78 187 DEU 0 0				
12103 1 5017 1	3.25 3.00 3.00	2 0 12.50 0	0 49.78 187 DEU 0 0				
12103 1 5017 1	4.25 4.00 4.00	2 0 10.60 0	0 49.78 187 DEU 0 0				
12103 1 5018 2	0.00 0.00 0.00	2 0 0.00 0	1 68.24 171 DEU 1 0				
12103 1 5018 2	0.25 0.00 0.00	1 1 0.00 200	1 68.24 171 DEU 0 0				
12103 1 5018 2	0.33 0.08 0.08	2 0 725.00 0	0 68.24 171 DEU 0 0				
12103 1 5018 2	0.50 0.25 0.25	2 0 596.00 0	0 68.24 171 DEU 0 0				
12103 1 5018 2	0.75 0.50 0.50	2 0 419.00 0	0 68.24 171 DEU 0 0				
12103 1 5018 2	1.25 1.00 1.00	2 0 227.00 0	0 68.24 171 DEU 0 0				
12103 1 5018 2	2.25 2.00 2.00	2 0 109.00 0	0 68.24 171 DEU 0 0				
12103 1 5018 2	3.25 3.00 3.00	2 0 54.00 0	0 68.24 171 DEU 0 0				
12103 1 5018 2	4.25 4.00 4.00	2 0 39.20 0	0 66.24 171 DEO U U				

Figure 4. Creating a new request is made through a web form.

 My account Log out 	Details Specifications Operational General Considerations Request Subject: *
SPECIFICATIONS	Current Modeler:
 ✓ Request • Add new 	521 user-id of the modeler.
 Track Module 	→ Therapeutic area(s)
 Add new Track 	→ Indication(s) → Purpose(s)
	─ ▶ Project(s)
	— ▶ Trial(s)
	Description:

Table 1. Quick overview of the differences between conventional statistical analysis and modeling and simulation at Novartis

	Conventional statistical analysis (at Novartis)	Modeling and simulation (at Novartis)
Data	Clinical endpoint and variable centric	Integrated
Specifications	 Extensive: Several report and analysis documents Several locations 	 Simplified: Web form Centralized: Database (this poster)
Pooling	Occasional	Routine
Scope	Mostly trial specific: clinical assessments (eg. safety, efficacy), organized by development stage (eg. Early vs full development)	Often integration of data (eg. Dose, drug concentrations, clinical read outs, efficacy or safety, etc), contribution throughout the whole drug lifecycle (from pre-clinical till post registration)

A minimal set of variables is required for any modeling datasets (eg. Unique identifier of individuals, elapsed time after first event, gender, etc...). The definition of these variables is standard and should not be changed. However, any additional variables may be added if allocated to one category and given required property details. Datasets then remain customizable beyond a well-defined and set of standard variables.

The building of modeling datasets has been decomposed into two parts: reusable modules and request specific (Figure 2).

Modules are generic programs that extract from the source data and preprocess them to a standardized format. There are as many module programs as there are types of clinical read outs. Their location is fixed and unique for a compound. Their format is kept as consistent as possible during the whole drug development lifecycle. Derivations and mapping operations are kept minimal in these programs to make them as reusable as possible in as many requests as possible. Module programs are executed from within a request program to prevent the creation of permanent datasets which could cause data access (e.g. interim analysis) and storage (e.g. space) issues. The way modules are called allow to select the studies to be pooled as they may not all be required for any requests.

Figure 2. The two steps of building a modeling data file



	A
Cours Downsont	
Save Request	

Once requests are entered, the user can navigate through all existing requests by choosing amongst different activity reports (Figure 5) by project, modeling purpose, therapeutic area or indication, owner and status of the request, etc... This tool offers different access permissions and dashboards depending on the role of the user. Three roles are currently implemented: modeler (with a focus on monitoring progress on requests and requesting new data), programmer (with a focus on more operational tasks) and manager (with a focus on monitoring the work load and drawing activity reports by project).

Figure 5. Default activity reports by type of requests

PINAUGR1	Reques	st Tracker sts by project						
- Log out	All R	equests All Open Requests	Ongoing Requests	Pending Requests	Completed R	equests Ca	ncelled Requests	My Requests
SPECIFICATIONS	Request #	Request Subject	Updated date 🔻	Currrent Modeler Curr	ent Programmer	Current Status	Validation Level	Edit link
✓ Request	25249	PopPKPD - covariate search	2011-04-20 04:26	bucl	nhvi1	Ongoing	Non-critical	edit
Add new	25243	PK modeling	2011-04-19 11:10	pina	ugr1	Ongoing	Most-critical	edit
 Track Module 	25247	Final analysis support FDP decision	2011-04-14 10:36	ette	ge1	Ongoing	Critical	edit
Woudie	05040	E and Distance in the second	2011-04-12-14-52	mad	lovhu1	Ongoing	Non critical	odit

The specification part has been designed to fit with the data structure appropriate to M&S. The online view describes the requested dataset in two different tables (Figure 6). The table of variables shows a description of each column, and the table of events provides a similar description for the dataset's rows. At any point in time, the user can display the specification table and / or export it to a separate document.

Figure 6. Description of the requested dataset

View Edit Re	visions	Clone						
e quest #: 25243 ersion #: 25533								
meline: 04-13-2011								
ed, 2011-04-13 05:25 — pinaugr1								
Details Specification	ns Ope	erational	General Considerations					
able of Variables: ategory	Name	e	Label		Unit	IME	Description	
dentification Variable	STUE	γC	Study Number			Study number		
dentification Variable	CENT	TER	Study Center					
dentification Variable	SUB	J	Subject identifier					
dentification Variable	ID		Unique subject ID					
dentification Variable	STYS	SID1A	Clinical Study Subject ID					
Event	CMT		Compartment				See table of	events
Event	EVID		Event ID				See table of	events
Event	AMT		Dose Amount				See table of	events
Event	LIDV		Linear dependent variable				See table of	events
Event	EVDI	г	Actual event date and time				See table of	events
īme Variable	TIME		Time since first event		h			
īme Variable	TIM2		Elapsed time since most recent administration		h		Administrati	on of ABC123
īme Variable	NT		Nominal time		h	Nominal time		
Fime Dependent Covariate	L2		Number of distinct event				Nonmem va - from 1 to n - 'distinct' / ' - sort by sty	iable. : +1 at each new event new' = different EVDT sid1a evdt evid pfldys1n.
Time Dependent Covariate	GRG		Numbering within repeated events				- from 1 to n - 'repeated' = - sort by sty	: +1 at each event with same EVDT = same EVDT sid1a evdt evid pfldys1n.
Time Dependent Covariate	000		Occasion variable				+1 at each r	iew dose level, should match the changes in DOSE
ïme Dependent Covariate	DOSI	Ε	Periodic dose level			Carry forward	[Provide dru	g and unit]
Time Independent Covariate	HTO		Height at baseline		cm		[Provide cod	e/decode - definition of the baseline]
Fime Independent Covariate	BMIO		Body mass index at baseline		kg/cm2	Body Mass Index	[Provide cod	e/decode - definition of the baseline]
Time Independent Covariate	WTO		Weight at baseline		kg		[Provide cod	e/decode - definition of the baseline]
Fime Independent Covariate	AGE	0	Age at demographic visit		Years			
Time Independent Covariate	RACE	Ξ	Race				[Provide cod	e/decode]
Time Independent Covariate	SEX		Sex				[Provide cod	e/decode]
a <mark>ble of Events:</mark> abel CMT	EVID AM	Т	EVDT	LIDV			Unit	Description
Drug administration 1	1 Dos	se Amoun	t [Source variables or reference to module]	0			[User input]	[Provide the drug to be administered in case there are several]
Pk concentrations 2	0 0		[Source variables or reference to module]	[Source va	ariables or	reference to module]	[User input]	[Provide the analyte, biological matrix, etc if applicable]

Objectives

Industrialize the data sourcing for model based drug development in a regulatory, consistent and standardized environment by:

- promoting reusability
- building institutional knowledge
- pooling data cumulatively
- moving towards standardized data structures

Methods

A new business model to generate the data has been developed:

- The Data Source Name (DSN) or composition of the input dataset has been redefined: simplifying and posing the vocabulary and grammar used to describe data required for any modeling activities (i.e. independently from the model or software to be used).
- The program organization has been adapted to enable a one-time extraction of data from each study, and an integration of the different exam data types together, in consideration of the regulatory environmental constraints, such as versioning and the access rights to the data.
- A Data Request Tracker (DRT) has been developed to track the modeling data requests, and to support a better management of the resources.

Results

A typical dataset structure includes two main variable types:

- Event: these cause the dataset to grow in number of records (rows).
- Covariates: these make the dataset grow in number of variables (columns).

These two variable types have been adapted to the most widely used software for non-linear mixed effect modeling called NONMEM® (Figure 1). This format can then be easily transposed to most types of modeling purpose and software. In this format, the following categories can be observed:

Request program integrates (Figure 3) the derived data from the module and convert them into any type of format before saving the data as a permanent output. They are unique and specific to the request in terms of content and location. A request program can call modules from several compounds and/ or indications, which data need to be integrated into a modeling data file.

Figure 3. Raw data integration over time is performed in the request program.

	Adı	ministration dataset			PK dataset					PD dataset				
CTR1N	SBJ1N	SMDDOS10	TDD1N	CTR1N	SBJ1N	SMP D T	CONC	CTR1N	SBJ1N	SMP D T	CONC			
			Total daily											
(Center	(Subject	Administration	dose	(Center	(Subject	Sampling	Concentration	(Center	(Subject	Sampling	Concentration			
number)	number)	date and time	amount mg	number)	number)	date and time	in ug/mL	number)	number)	date and time	in ug/mL			
1	5101	01JAN2000:09:00	150	1	5101	01JAN2000:09:10	21	1	5101	01JAN2000:09:10	21			
1	5101	02JAN2000:10:00	150	1	5101	01JAN2000:09:20	100	1	5101	01JAN2000:09:20	100			
1	5101	03JAN2000:09:00	150	1	5101	01JAN2000:09:30	90	1	5101	01JAN2000:09:30	90			
3	5102	01FEB2000:08:00	150	1	5101	03JAN2000:09:10	40	1	5101	03JAN2000:09:10	40			
3	5102	02FEB2000:09:00	150	1	5101	03JAN2000:09:20	150	1	5101	03JAN2000:09:20	150			
3	5102	03FEB2000:09:00	150	3	5102	01FEB2000:08:10	40	3	5102	01FEB2000:08:10	40			
				3	5102	01FEB2000:08:40	130	3	5102	01FEB2000:08:40	130			
				3	5102	01FEB2000:09:10	120	3	5102	01FEB2000:09:10	120			
				3	5102	03FEB2000:09:10	35	3	5102	03FEB2000:09:10	35			
				12			234 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142 - 142	6			101			

Integration of raw data over time

D	CENT	SUB	CMT	TIME	DV	EVID	DOSE	AMT	MDV	EVDT	
Derived ID	Center number	Subject number	1:Dose 2:PK 3:PD	Time since first event (h)	Dependent variable	Event identifier	Dose in mg	Daily dose amount	Missing dependent variable	Actual date and	d time
1	1	5101	1	0	0	1	150	150	1	01JAN2000:09	:00
1	1	5101	2	0.1667	21	0	150	0	0	01JAN2000:09	9:10
1	1	5101	3	0.1667	31	0	150	0	0	01JAN2000:09	9:10
1	1	5101	2	0.3333	100	0	150	0	0	01JAN2000:09	9:20
1	1	5101	3	0.3333	110	0	150	0	0	01JAN2000:09	9:20
1	1	5101	2	0.5	90	0	150	0	0	01JAN2000:09	9:30
1	1	5101	3	0.5	10	0	150	0	0	01JAN2000:09	9:30
1	1	5101	1	25	0	1	150	150	1	02JAN2000:10	0:00
1		5101	1	48	0	1	150	150	1	03JAN2000:09	9:00
1	1	5101	2	48.1667	40	0	150	0	0	03JAN2000:09	9:10
1	1	5101	3	48.1667	45	0	150	0	0	03JAN2000:09	9:10
1	1	5101	2	48.3333	150	0	150	0	0	03JAN2000:09	9:20
1	1	5101	3	48.3333	153	0	150	0	0	03JAN2000:09	9:20
2	3	5102	1	0	0	1	200	200	1	01FEB2000:08	8:00
2	3	5102	2	0.1667	40	0	200	0	0	01FEB2000:08	3:10
2	3	5102	3	0.1667	41	0	200	0	0	01FEB2000:08	8:10
2	3	5102	2	0.6667	130	0	200	0	0	01FEB2000:08	8:40
2	3	5102	3	0.6667	128	0	200	0	0	01FEB2000:08	8:40
2	3	5102	2	0.8333	0	0	200	0	1	02FEB2000:08	3:50
2	3	5102	3	0.8333	0	0	200	0	1	02FEB2000:08	3:50
2	3	5102	1	25	0	1	200	200	1	02FEB2000:09	9:00
2	3	5102	1	49	0	1	150	150	1	03FEB2000:09	9:00
2	3	5102	2	49.1667	35	0	150	0	0	03FEB2000:09	9:10
2	3	5102	3	49.1667	32	0	150	0	0	03FEB2000:09	9:10

Conclusions

M&S data preparation requires constant data integration from a variety of raw data sources throughout the project lifecycle. A good understanding of our Data Source Name (DSN) and how this data should be integrated is the first step toward standardization and improved input data quality to model based analysis⁵. The optimized and consistent workflow and program organization is supporting our productivity efforts. A dedicated tracking tool enables proper documentation and recording of every new request. This new business model presents the following advantages:

- Allows custom datasets to be built in a standard frame, ensuring a minimal fixed core set of variables and definitions is used.
- Enables the integration of raw data over time and promote reusability by modules.
- Optimizes a consistent workflow and program organization promoting productivity.
- Records every single data request enabling institutional knowledge to be built and promoting reusability of existing work accomplished.

- **Identification variables**: to what study and individual do the records belong on the current row? E.g. ID is the unique identifier of each individual contributing to the model, STUDY is the numeric value of the study the individual is coming from, etc...
- **Time variables**: when the result has been measured relative to a certain point of reference? E.g. TIME is the elapsed time from the very first event; TAD is the elapsed time from the most recent drug administration, etc...
- **Event variables**: events are composed of administrations (what is given to the subject), observations (measurements taken from the subject), or other (imputed records for richer simulations). They are coded in the dataset with a minimum of four variables: CMT (a code to identify the type of event), EVID (the event identifier: administration or observation), LIDV (the result of an observation), AMT (the amount of drug administered).

The data requests, needed to support modeling activities, can now be requested through a centralized "Data Request Tracker" platform replacing a previous Excel based process.

This tool is web-based to enable a rapid access from any computer in the Novartis network (underlying technologies: DRUPAL⁴, JQuery, Ajax). It requires user authentication and operates over a relational database which allows for real time activity reporting. Users can enter information about their modeling activity and the requested data by navigating through several structured tabs: Details, Specification, Operational, and General Considerations (Figure 4).

• Simplified and detailed, systematic and centralized documentation on the process and specifications to generate the data used for model based analysis facilitating generation of documents for Health Authorities.

The benefits of this new business model are:

- Faster and well documented access to data,
- Better quality of modeling data files through development of standards and routines,
- Extension across compounds and across indications model based drug development.

Acknowledgements

The authors would like to thank Jean-Louis Steimer and the programmers of M&S for their continuous support and valuable insights included in this poster.

This initiative was supported by Novartis Pharma AG, Basel, Switzerland. Copyright © 2011 Novartis Pharma AG, Basel, Switzerland. All rights reserved.

References

- 1. Pinault G., et al. Quality, Efficiency and Industrialization Initiatives during the evolution of a dedicated SAS Programming Group. PAGE 18 (2009) [Abstract 1513]. Available at http://www.page-meeting. org/?abstract=1513.
- 2. Joga Gobburu, "Lean & Apply" Paradigm in Drug Development: How disease-drug-trial models can aid decisions. Available at: http://www.rosaandco.com/webinarGobburu.html
- 3. Shafi C. Creating NONMEM datasets how to escape the nightmare. Pharmaceutical programming (2010). Vol.3, No.2, p.80-83.
- 4. DRUPAL organization (www.drupal.org)
- 5. Buchheit V., et al. Efficient quality review for modeling input dataset. PAGE 20 (2011) [Abstract 2041]. Available at http://www.page-meeting.org/default.asp?abstract=2041.

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