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

NOVARTIS

Background

The value of the Modeling and Simulation (M&S) group has been recognized within the organization[1]. Development of new drugs now requires more systematic and more frequent model based insights from M&S[2]. 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 challenges[3]. 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.

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)

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:

- **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).

- **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*


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 pre-process 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*
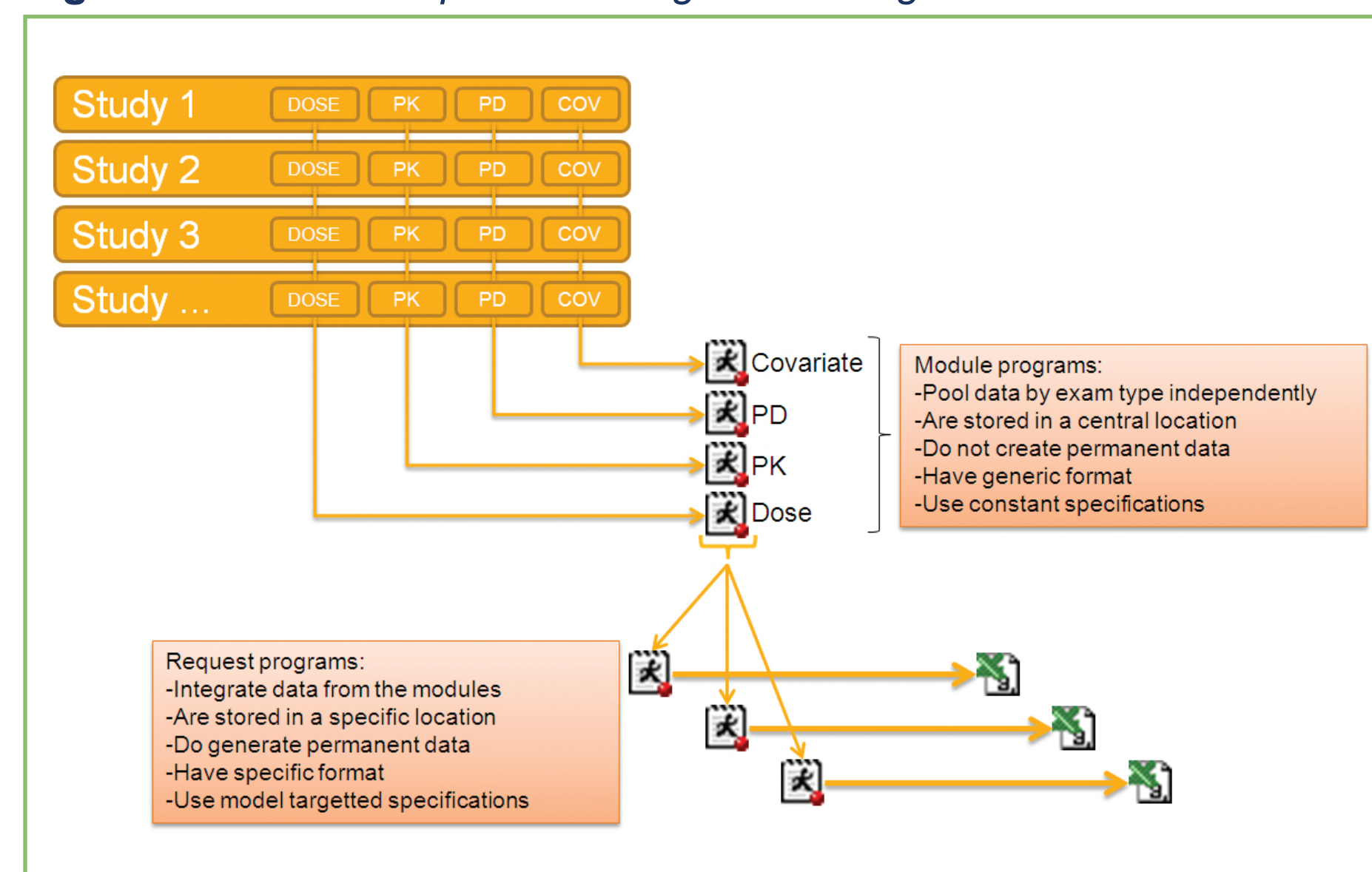
- **Request** program integrates **(Figure 3)** the derived data from the module and convert them into any type of format before saving the data as a permanent output. They are unique and specific to the request in terms of content and location. A request program can call modules from several compounds 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.*

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: DRUPAL[4], 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)**.

Figure 4. *Creating a new request is made through a web form.*
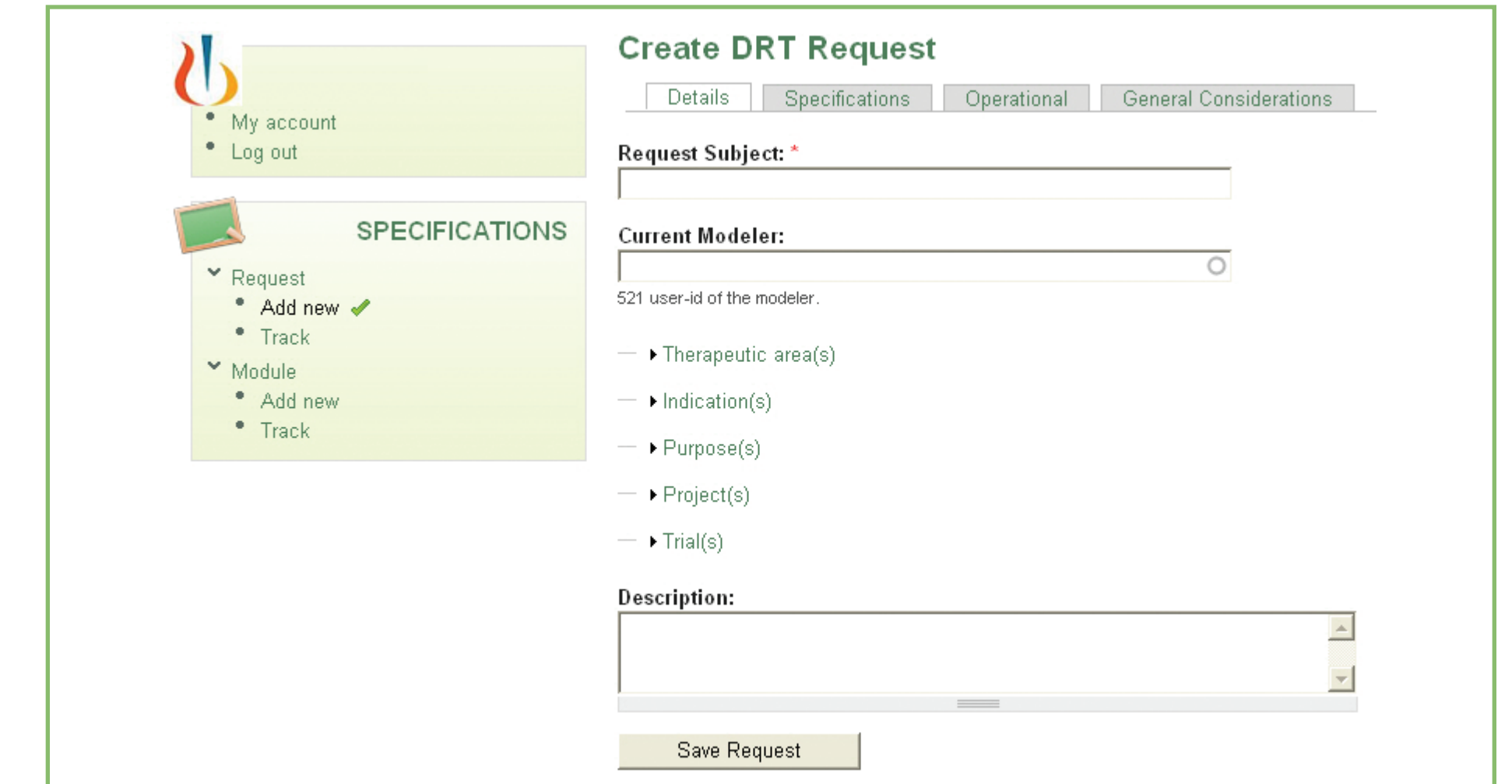
Once requests are entered, the user can navigate through all existing requests by choosing amongst different activity reports **(Figure 5)** by project, modeling purpose, therapeutic area or indication, owner and status of the request, etc... This tool offers different access permissions and dashboards 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*

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*

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 analysis[5]. 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.
- 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

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-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.