

Maintaining consistency between a substance database and single study databases

Dieter Hartz, Boehringer Ingelheim, Ingelheim, Germany

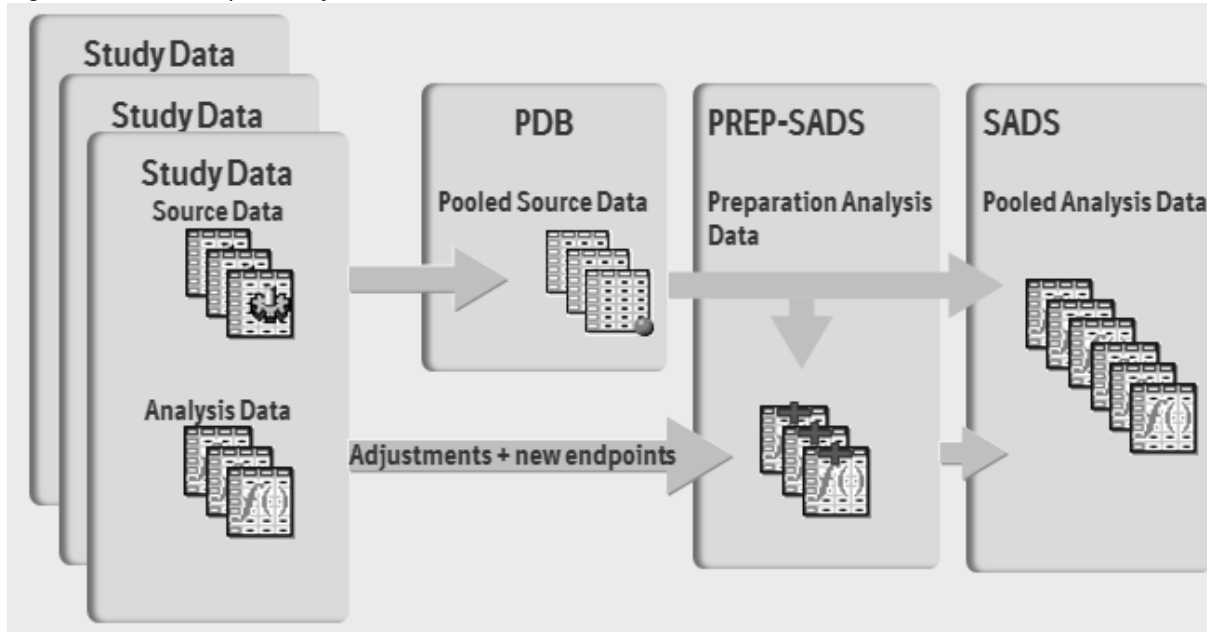
ABSTRACT

Regulatory submissions, ad hoc requests and publications require pooled analyses of study data. For this purpose we have established a substance database for a DPP-4 inhibitor against type II diabetes. Various challenges arise when the database is being updated with either new studies or endpoints. On source data level the integrity of project data base (PDB) from new data lock point has to be maintained with previous PDB. The same holds true on substance analysis dataset (SADS) level where contents and structure of new study data have to be integrated. Hardcodings done on study level also have to be considered on substance level. Algorithm changes will lead to global SADS update while study algorithms should be maintained as well. These and other challenges and possible solutions will be discussed.

INTRODUCTION AND OBJECTIVES

- In order to allow for pooled analyses of study data for regulatory submissions, ad hoc requests and publications a substance database for a DPP-4 inhibitor against type II diabetes was established.
- A snapshot of this database consists of substance analysis datasets (SADS) and source datasets at the appropriate data lock points in a project database (PDB).
- Only completed studies or data from studies locked for planned interim analysis are contained in the substance database. About 30 studies are contained in the current snapshot.
- As needed, the database is being expanded with data of newly completed studies and stored as a new snapshot. If new endpoints are required these are being derived across studies.
- Each snapshot is used for a defined set of analysis tasks. While analyses are performed from last snapshot the next snapshot can already be worked on. Thus continuous work on the substance database is possible in order to avoid the work load of pooling the study data just before each submission or any other analysis package.
- For reproducibility of previous analyses the consistency between different snapshots as well as between study data and snapshot data should be maintained.

Figure 1: Relationship of study data, PDB, and SADS.



MAIN CHALLENGES AND SOLUTIONS

CONCEPT OF DERIVING SAFETY AND EFFICACY ANALYSIS DATA

- The consistency of **safety data** largely depends on the consistency of PDB datasets. Analyses are directly done from PDB source datasets (AE, LAB) using BI internal standard macros or from SADS datasets newly derived from PDB source datasets for each snapshot (SADS containing treatment information, rescue information, analysis populations). Therefore, new snapshot PDB datasets are compared with previous snapshot PDB datasets and new snapshot SADS datasets are compared with previous snapshot SADS datasets for consistency with previously existing data.
- The consistency of **efficacy data** largely depends on study analysis datasets because efficacy endpoints are directly taken from study efficacy datasets without re-derivation of study endpoints. Only new endpoints will be added or new imputation methods will be implemented in each snapshot. New studies to be integrated are checked for efficacy variables and attributes (name, length, format, label), variable codes and decodes, categorisations of endpoints and their decodes, units of endpoints concerned, and missing values of variable for endpoint value or endpoint date. The finding of inconsistencies will prompt adjustments to meet the standards of SADS efficacy data. The study efficacy analysis dataset is converted into a study preparation efficacy analysis dataset for pooling. New snapshot SADS datasets are compared against previous snapshot SADS datasets for consistency with previously existing data.

INTRODUCTION OF NEW ALGORITHMS FOR EXISTING ENDPOINTS DURING PROJECT DEVELOPMENT – AN EXAMPLE OF EGFR (= ESTIMATED GLOMERULAR FILTRATION RATE)

- The original MDRD (= Modification of Diet in Renal Disease) formula should only apply for patients in studies in which the method of creatinine measurement was not calibrated to be IDMS (= Isotope Dilution Mass Spectrometry) traceable:

$$\text{eGFR (mL/min/1.73m}^2\text{)} = 186 \times \text{creatinine (mg/dL)}^{-1.154} \times \text{age}^{-0.203} \times (0.742 \text{ if female}) \times (1.212 \text{ if black})$$
- A new formula should be applied for patients in studies in which the method of creatinine measurement was calibrated to be IDMS traceable:

$$\text{eGFR (mL/min/1.73m}^2\text{)} = 175 \times \text{creatinine (mg/dL)}^{-1.154} \times \text{age}^{-0.203} \times (0.742 \text{ if female}) \times (1.212 \text{ if black})$$
- For Japanese patients yet another new formula was introduced:

$$\text{eGFR (mL/min/1.73m}^2\text{)} = 194 \times \text{creatinine (mg/dL)}^{-1.094} \times \text{age}^{-0.287} \times (0.739 \text{ if female}).$$

To maintain reproducibility of study results for eGFR the different formulas were assigned to different parameter names (EGFR using original MDRD formula, EGFRJ using new formula, EGFRJ using Japanese formula). For studies that had originally used only EGFR also EGFRJ was derived because for the majority of them the new formula was to be applied. For Japanese patients also EGFRJ was derived. The selection which formula per patient is used was then collected in a new parameter (EGFRJ). A disadvantage of this approach is the expansion of the

PhUSE 2013

SADS datasets.

MANAGING HARDCODINGS

Hardcodings are sometimes employed in study data to exclude patient data or to correct obviously inconsistent data in a meaningful way. If such hardcodings were employed for study ADS they should be maintained for SADS as well.

- A hardcoding tool has been developed that allows for better tracing of hardcodings and demands the approval by the responsible statistician before it can be employed.
- Any hardcoding has to be entered into an EXCEL sheet with the SAS® code for the hardcoding and a corresponding ID for approval.
- The data of the EXCEL sheet are exported into a SAS dataset.
- Within the SADS program that employs the hardcoding the ID is being invoked as positional parameter within the hardcoding macro call and only if the hardcoding was approved by the statistician the program will run without giving an ERROR message from the macro call.

The employment of the hardcoding tool allows for easy documentation and traceability of the hardcodings.

FURTHER CHALLENGES AND SOLUTIONS

Challenge	Solution
➤ The periodic update of dictionaries for adverse events, medical history and therapies (MedDRA, WHO-DD) leads to slight inconsistencies between PDB datasets for each snapshot with respect to coded information while investigator verbatim text should remain the same.	➤ PDB safety datasets also contain LLT coding information from previous and originally used MedDRA/WHO-DD versions so that original outputs could be reproduced.
➤ Changes in eCRFs might lead to new variables in PDB source dataset while other variables might become obsolete.	➤ Variables becoming obsolete in PDB source datasets were renamed with an underscore so that they will not be used by accident but can be used to compare with previous content of PDB snapshot datasets. Where possible contents of obsolete variables were re-mapped to the new variables.
➤ Introduction of new standard units for laboratory parameters leads to change of values for parameters concerned.	➤ In PDB the SI standard units from new study data were converted back into conventional units so that original outputs from previous snapshot could be reproduced.
➤ The request of additional imputation methods for efficacy endpoints might also demand for the re-derivation of already existing efficacy endpoints.	➤ When re-derivation of existing imputations was necessary baseline selection rules employed in the studies were also employed in SADS. Deviating time windows were adjusted to project standards.
➤ Derivation of new efficacy endpoints from parameters also reported in safety analyses.	➤ Whenever possible data were crosschecked with data derived for safety analyses, especially for consistency in selected baseline values.

CONCLUSIONS

The maintenance of consistency between a pooled substance database and study data supports the long term goal of reproducibility of previously reported analyses from newly pooled data. This was mainly achieved by maintaining the endpoints from the single studies while in addition deriving new endpoints for analyses across studies that employ different algorithms. As a consequence there is a multiplicity of endpoints in the substance analysis data sets. In order to avoid working with large datasets for submissions a subset of the SADS can be generated from a defined snapshot. These project analysis datasets then serve as a base for an analysis package. The concept of building defined snapshots of the substance database for reporting also demands consistency checking between snapshots. While at first this appears to be extra work, it allows the easy detection of unexpected changes so that erroneous reporting is avoided. Furthermore it is possible to prepare analyses from one snapshot while working at the same time on the generation of the next snapshot.

ACKNOWLEDGMENTS

Michael Knoessl (BI), Felix Mader (BI), André Haugg(BI), Ralf Minkenberg(BI), Barbara Doll (BI)

CONTACT INFORMATION

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

Dieter Hartz
Boehringer Ingelheim Pharma GmbH & Co. KG
Binger Strasse 173
55216 Ingelheim am Rhein
Work Phone: +49 61 32 77 14 25 00
Email: dieter.hartz@boehringer-ingelheim.com

Brand and product names are trademarks of their respective companies.