Visions of a Drug Development Data Warehouse

Paul Vervuren, Organon, Oss, The Netherlands

ABSTRACT

Data warehouses are information systems designed to deliver information that cannot be provided by operational systems. At Organon a data warehouse project has been initiated with the primary aims to allow study data exploration across compounds, to link clinical, pre-clinical and research data, and make drug development data more easily and widely available. This paper presents some of the experiences and results of an early phase of a data warehouse project, which was restricted to clinical data. It includes a proposal for the simultaneous implementation of a data warehouse and the CDISC SDTM standards, and a query/reporting application build in SAS/AF® in support of a data warehouse pilot.

INTRODUCTION

The goal of the processing and analysis of study data within pharmaceutical companies is to provide information for regulatory authorities, the scientific community and treating physicians in order to introduce and successfully apply new drugs. Typically, the organization and standardization of operational processes is aimed at the compound level, allowing easy data integration for submission but making integration of data across compounds much more difficult. Driven by the awareness that potential knowledge within our company may thus be unused, and encouraged by external developments, in particular the CDISC (Clinical Data Interchange Standards Consortium) data standards, a data warehouse project was initiated at the Biometrics department of Organon.

In this paper I will present some of the experiences and results of an early phase of the data warehouse project. Ultimately, this project aims to include data from research to clinical development, but it started off with clinical data; the data we are most familiar with in the Biometrics department. After giving an outline of the project, I will (1) describe how we envision the simultaneous implementation of a data warehouse and the CDISC SDTM standards within our operational systems and processes; and (2) present a query/reporting application that was built for demonstration purposes within a pilot. The reasons for building a data warehouse at Organon, as expressed below, are very similar to those described by other pharmaceutical companies (e.g. Koprowski 2000). The approaches and solutions will vary to some extent between organizations depending on the systems in place and company characteristics like culture, size, structure and expertise.

OUTLINE OF THE PROJECT

In a start-up phase of the data warehouse project interviews were held with major stakeholders in our organization (representing departments ranging from Research to Medical Affairs). These interviews emphasized the need for easy access to data to both biometricians and non-biometricians, the ability to conduct cross-compound analyses, and the desire to combine pre-clinical and clinical data. The benefits of a facility that could provide this would be:

Increased and wider availability of knowledge

- Data exploration and mining across compounds
- Linkage of clinical, pre-clinical and research data
- Design of new trials
- Business Intelligence

Improved services

- User-friendly system (accessible to ‘everyone’)
- Faster and easier processing of ad-hoc requests (e.g. reports/analyses for publications or questions from regulatory authorities)
- Provide data for web publication (Clinical Trial Register)
- Create data structures for regulatory submission

Based on this information, a next phase of the data warehouse project, referred to as the 1st increment, started early 2005.

OBJECTIVES

The objective of the project's 1st increment was to create a roadmap for the implementation of a data warehouse. This encompassed:

- The review and selection of ETL solutions and front-end tools.
- A pilot to get real-life experience.
- Delivering a blueprint for maintenance and security.
- Creating awareness and obtaining buy in for the project within the company.

GUIDING PRINCIPLES

The guiding principles of the data warehouse project are:

- *A business-oriented approach*. The data warehouse can only be a success if it stores data in a manner that supports user demands. Knowledge of the data and its users are therefore crucial. (See Brown 2004.)
- *Open mindedness*. Innovation requires the willingness to break with existing habits and standards (biometricians are not used to giving free access to the data they 'own').
- *Use existing knowledge and expertise*. We own substantial in-house experience with SAS® and the development of standards. In-house data warehouse expertise was obtained through consultancy and training.
- *Incremental development*. Small steps force us to rapidly produce (preliminary) results, and to evaluate solutions, project directions and project management on a regular basis.
- *Attune to related external and internal developments*. We identified the CDISC STDM and ADaM standards, the FDA's JANUS data warehouse and the ICH E3 reporting guidance as important external developments. Internally, the Clinical Trial Registry (as published on the Organon website) was among the initiatives considered relevant to our project.

IMPACT ON OPERATIONAL SYSTEMS AND PROCEDURES

In this section the clinical data systems and related operational procedures at Organon will be introduced. After this, a possible strategy for data warehouse implementation will be presented, which encompasses the simultaneous introduction of the CDISC SDTM standards.

The main clinical data systems at Organon and their data flow are as follows:

1. Clinical data sources (data entry, e-CRFs, IVRS, spontaneous SAEs) are loaded in an Oracle Clinical database management system (DBMS)
2. Within the DBMS data is cleaned and checked for integrity and coding is provided for adverse events and medication.
3. Study data is extracted from the DBMS and stored as SAS data sets to be accessed by the Biometrics department.
4. Standard packages, consisting of SAS macros, as well as one-off SAS programs are used to create so called Master Data Sets (MDS). These data sets contain all raw data, derived variables (including analysis variables) and a number of key demographics and disposition variables, for analysis purposes.
5. SAS programs (using standard packages for common domains) use MDSs to create statistical output files and data listings, which are converted to PDF (Jansen 2001) or any other format to be included in study reports, publications etc.

Most relevant to the data warehouse is the Biometrics environment. The Biometrics work environment represents a globally standardized and controlled environment. It provides sufficient flexibility for ad-hoc analyses and allows the incorporation of data that does not comply with Organon/Biometrics standards, e.g. data from trials conducted in collaboration with other companies. Data standardization is achieved through global standards for the definition of key variables and compound level standards known as Master Data Set guidelines. Process control is achieved through:

- A so-called "Final Run" standard operating procedure. In this procedure all SAS programs, both for the generation of MDSs and output are sequentially submitted in batch mode.
- A SAS macro named MDU, which documents an executed Final Run and checks for consistency of the run, redundant programs, etc. See PhUSE 2005 presentation AS04 (Ament 2005).

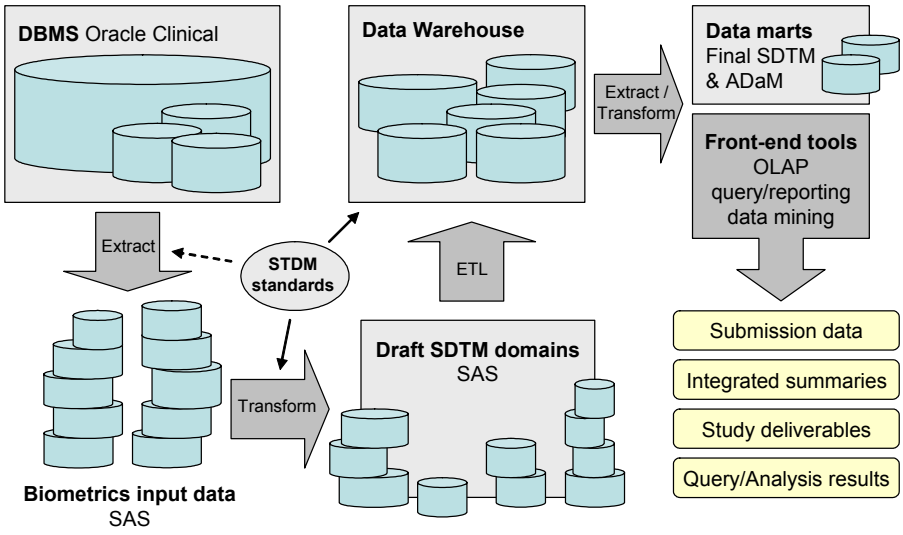

Figure 1 Envisioned implementation of SDTM and data warehouse.

IMPLEMENTATION OF SDTM STANDARDS AND DATA WAREHOUSE

CDISC is a consortium whose mission is "to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare" (www.cdisc.org/about/). There are several CDISC standards, of which the SDTM (Submission Data Tabulation Model) and ADaM (Analysis Dataset Model) are relevant within the framework of the data warehouse project. The SDTM model (CDISC/SDTM-IG 2004) describes the organization of all data collected during clinical trials ('CRF data'), but also includes study design information and protocol essentials. ADaM describes the organization of data used in statistical analysis together with related metadata. In deriving the analysis data from CRF data, the data is usually transformed. Purpose of the metadata (and other documentation) is to "provide clear and concise communication of the analytic results of a clinical trial from the sponsor to the regulatory reviewers, including statistical methods, transformations, assumptions derivations and imputations performed" (CDISC/ADaM 2004). The FDA has accepted the SDTM model as a submission standard in 2004. Just like the FDA expects SDTM standards to support their JANUS data warehouse (FDA/CDISC 2005) we expect it to support the implementation of our data warehouse through enforcement of standards.

Possible approaches for implementing the SDTM in our organization are (1) converting Master Data Sets (essentially a combination of SDTM and ADaM type data) to SDTM just before submission, or (2) replacing MDSs by draft SDTM compliant data sets. The advantage of the first approach is that it would leave current procedures largely unaffected, but there are obvious and large disadvantages, the most important being:

- Additional efforts will be required to transform the data into STDM
- Transformation into SDTM would be completely on the critical time path for submission
- Only submission-ready data would be available for loading into the data warehouse

Probably, the second approach would be a better solution.

The envisioned future data flow that incorporates SDTM and a data warehouse is outlined in figure 1. Compared to the present procedure, the step of extracting data from Oracle Clinical remains largely unchanged. However, some of the "general assumptions for all domains" (CDISC/SDTM-IG 2004) could be implemented at this level. For instance, transformation of raw dates: Oracle Clinical has format YYYYMMDD while the desired SDTM standard (ISO 8601) is YYYY-MM-DD. Actually, it would be desirable to have such standards implemented at the Clinical DBMS level (see Kenny & Litzsinger 2005, Shostak 2005). Commercial software to convert Oracle Clinical data to SDTM format is available (Kenny & Litzsinger 2005).

Most of the SDTM standards in this proposed scheme will be implemented in SAS programs that transform Biometrics input to (draft) SDTM domains. Its flexibility makes SAS a suitable tool for such

transformations. Examples of SAS macros for SDTM conversion can be found in Shostak (2005). The SDTM domain data sets being draft allows the presence of analysis variables and structures that are conducive to statistical analysis (similar to current Organon MDSs). In the envisioned situation these data sets represent the 'staging area' for the data warehouse. During the ETL (Extract Transform and Load) process between Draft SDTM domains and the data warehouse the data domains can be validated against CDISC standards.

Finally, extracts from the data warehouse (data marts) can be made to provide data for submission and ad-hoc queries or analyses using front-end tools. At this step the final SDTM and the ADaM data sets can be created for submission. This conforms to the 'hybrid method' for SDTM and ADaM development Kenny & Litzsinger (2005) proposed as a long-term solution.

We are still at the very start of implementing a data warehouse and CDISC standards and many aspects of the proposed scheme still need to be clarified. For example, to what extent SDTM standards can be implemented at the Oracle Clinical level within our setting, and where and how (draft) ADaM data sets and metadata will be created. It should also be noted that CDISC standards are still under development, in particular ADaM.

PILOT

As part of the 1st increment of the data warehouse project a pilot was conducted. The clinical data of an existing product served as pilot data (>140 studies, >13,000 subjects). This data partly consisted of legacy data, which was mapped to MDS standards. Shaping and integrating this data was a separate project and served a purpose of its own. Advantage of using this data in particular was that (1) the data was alive within and outside the biometrics department, (2) there were related business questions that could be used as real-life examples, and (3) the total size was considerable (making it awkward to use conventional browsing/query tools: an alternative solution could make a difference).

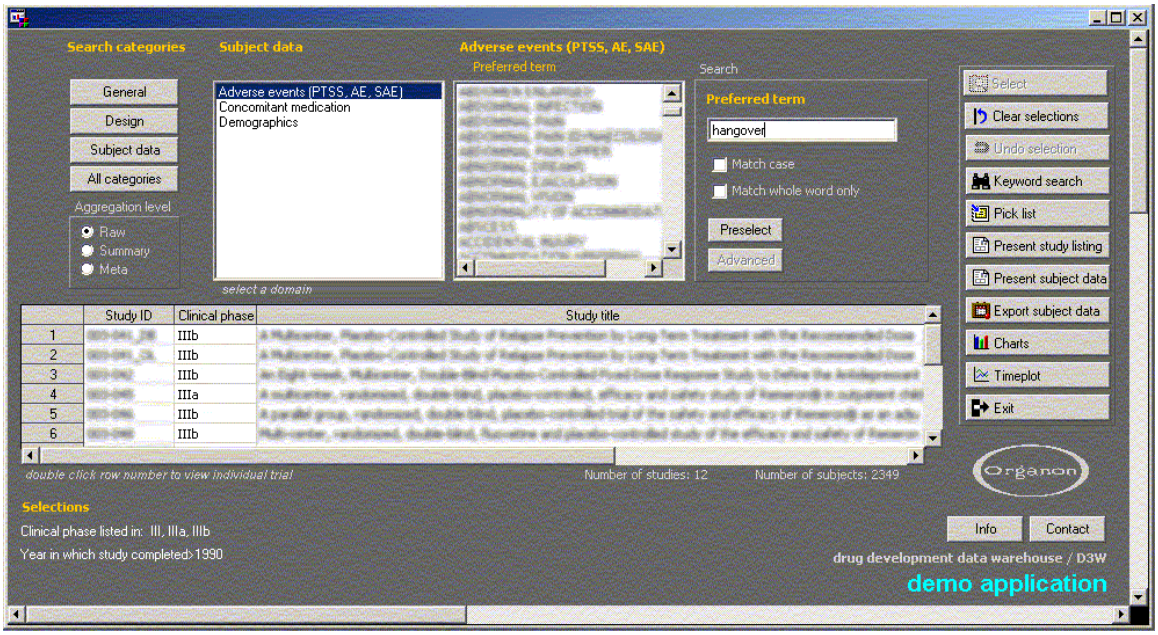

Figure 2 Main interface of the application.

In support of this pilot an application was developed using SAS/AF software (figure 2). This application very well suited the scope of this pilot but will not be used as a tool for the future data warehouse. The application appeared to be a useful means to:

- Visualize and clarify data warehouse concepts. Presentations of the application improved the level of understanding among potential users, resulting in more useful feedback.
- Obtain buy in for the project within the company.
- Define the basic requirements of the data warehouse and associated tools.

The application provides easy access to the integrated MDSs and contains query functions for numeric, codelist and free text variables. Users are allowed to select variables from different domains in subsequent queries; the results are combined at the subject level. Data thus selected can be exported or used to produce data listings and graphs. Examples of business questions used to design and test the application were:

1. How many subjects aged xx-xx had event xx in studies of a planned duration up to xx weeks and containing at least xx treated subjects.
2. Which studies used efficacy parameter xx.
3. Were there any trends in the efficacy variable xx over time when comparing treatment groups xx and xx, and was this different for men and women.
4. How many subjects were involved in phase I studies
5. Which studies had a crossover design (based on information in title/objectives)

DATA AND META DATA

Besides subject data, stored in separate domain data sets (Demographics, Vital Signs, Concomitant Medication, Adverse Events, etc.) additional data was needed for the application (figure 3). To resolve questions like 4 and 5 it was necessary to provide general trial information and design features (Trial Summary). This information was partly taken from study reports (e.g. titles and objectives) and entered manually, and partly derived from the integrated MDSs. Summarized data (for example, frequency counts of the number of treated subjects), needed to answer question like the 1st in above list, was derived from the integrated MDSs. Moreover, the application requires meta data in order to convert the variable label selected by a user to a variable name, access the proper domain data set to select the variable, and behave in a manner appropriate for the value of the variable. Different types of meta data are present in the application to accommodate this. The following table gives an example of meta data that describes some Trial Summary variables:

Label	Variable	Type	Aggregation level	Category	Selection mode
Clinical phase	CLPHASE	CHAR	RAW	GENERAL	MULTIPLE SELECTIONS
Planned treatment duration	PLTRDUR	NUM	RAW	GENERAL	NA
Title	TITLE	CHAR	RAW	GENERAL	KEYWORD SEARCH

Some of these attributes were manually assigned, for example ‘Aggregation level’ (to distinguish between original and summarized data), ‘Category’ (used to group items in a listbox) and ‘Selection mode’ (controls the type and behavior of the field displayed when the item is selected). Another type of meta data underlying the application are data sets that contain unique values of codelist variables and coded terms.

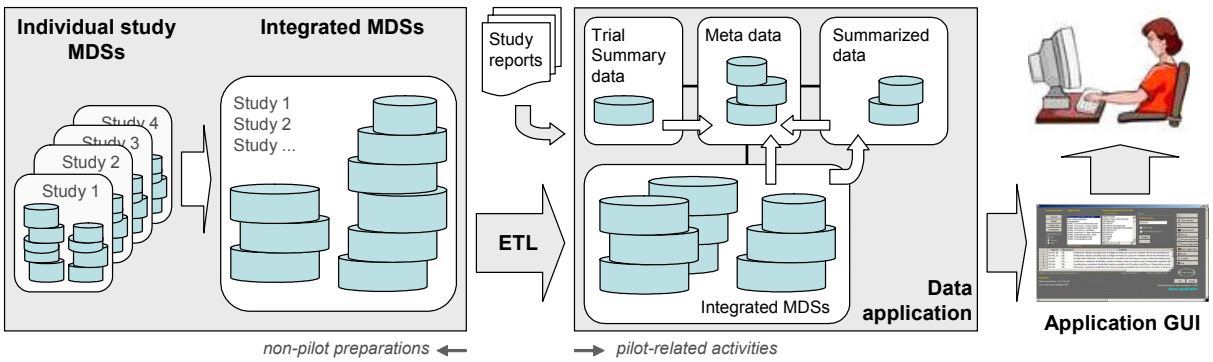

Figure 3 Flowchart of data used by the application.

During the ETL step (figure 3) some simple transformations and selections were conducted. For instance only studies having an entry in the Trial Summary data set were selected (this means general information must be available: phase, title, objectives). Initially these steps were executed by Base SAS programs. An evaluation of possible ETL future tools (one of the objectives of the project) resulted in the selection of SAS ETL Studio as a possible candidate. We are currently testing SAS ETL Studio, using it to set up the integrated MDSs and meta data for the application.

CONCLUSION

After the 1st increment of the data warehouse project we are still at the very early stage of building a drug development data warehouse. Yet we feel we have been making substantial progress within the limited time frame of this project. Some of the results and experiences of this projects have been presented in this paper. In short, the accomplishments of the project are:

- A road map for the implementation of a data warehouse (restricted initially to clinical data) has been developed, which involves the simultaneous implementation of the CDISC SDTM standards.
- A pilot was completed that was successful in:
  - Generating and testing business questions
  - Establishing the need for Trial Summary data
  - Creating awareness and understanding of the project
- SAS ETL Studio has been selected as a candidate ETL tool for the data warehouse.

Factors that presumably contributed to these accomplishments are a business-oriented approach and incremental development (as outlined in our guiding principles). The next steps in our project are:

- Setting up data structures for the data warehouse and loading data
- Hardware and infrastructure selection
- Define user-groups and estimate user numbers
- Select front-end tool(s) for the different user-groups
- Adapt existing working procedures
- Define staff

REFERENCES

Ament, H. (2005). "Macro MDU: a flowchart of used programs, datasets and TLF's based upon the Final Run", *Proceedings of the First Conference of the Pharmaceutical Users Software Exchange (PhUSE)*.

Brown, M. (2004). "8 Characteristics of a Successful Data Warehouse". *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference*.

CDISC/ADaM (2004). "Statistical Analysis Dataset Model: General Considerations Version 1.0"

CDISC/STDM-IG (2004). "Study Data Tabulation Model Implementation Guide: Human Clinical Trials", Document version 1.01. Submission Data Standards Version 3.1.

FDA/CDISC (2005). "Walking Down the Critical Path: The Application of Data Standards to FDA Submission. A Discussion Paper by CDISC and FDA".

Jansen, L. (2001). "Creating PDF documents including hyper-links, bookmarks and a Table of Contents with the SAS® software", *Proceedings of the Annual Conference of the Pharmaceutical Industry SAS® Users Group (PharmaSUG)*.

Kenny, S.J. & M.A. Litzsinger (2005). "Strategies for Implementing SDTM and ADaM Standards", *Proceedings of the Annual Conference of the Pharmaceutical Industry SAS® Users Group (PharmaSUG)*.

Koprowski S.P. Jr., Barrett J.S. & D.J. Fowler (2000). "Constructing a Data Warehouse for Pharmacokinetic Data", *Proceedings of the Annual Conference of the Pharmaceutical Industry SAS® Users Group (PharmaSUG)*.

Shostack, J. (2005). "Implementation of the CDISC SDTM at the Duke Clinical Research Institute", *Proceedings of the Annual Conference of the Pharmaceutical Industry SAS® Users Group (PharmaSUG)*.

ACKNOWLEDGMENTS

The author would like to thank the members of the drug development data warehouse (D_3W) project team whose ideas and efforts are reflected in this paper. Useful suggestions regarding the implementation of SDTM standards and the data warehouse were given by Lex Jansen. The manuscript was kindly reviewed by Kit Roes, and the D_3W project leaders Egbert Biesheuvel and Tom Peters.

CONTACT INFORMATION

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

Paul Vervuren
Biometrics/Clinical Information department
NV Organon
PO Box 20
5340BH Oss
Work Phone: +31 412 663921
Fax: +31 412 662516
Email: Paul.Vervuren@organon.com
Web: www.organon.com