Paper 117-25

Constructing a Data Warehouse for Pharmacokinetic Data

S.P. Koprowski, Jr., J.S. Barrett, DuPont Pharmaceuticals Company and D.J. Fowler, SAS Institute, Inc.

ABSTRACT

The construction of a warehouse using data from DuPont Pharmaceutical's Clinical Pharmacokinetics Group with the SAS/Warehouse Administrator™ product is described herein. Data sources incorporated into the pharmacokinetic warehouse existed in multiple formats (Oracle, Excel, SAS data sets, etc) and were represented by numerous individual files, rather then few large files. In most cases, the data density enjoys low granularity. The pharmacokinetic data warehouse draws data from many sources represented by multiple departments and spans long drug development cycles with varying degrees of detail. Additionally, data originated from both internal legacy systems or was obtained from an external contract research organization (CRO). Remote systems have been easily identified within the warehouse and views rather than distinct copies have been incorporated as data sources. The SAS/Warehouse Administrator™ product was essential to this effort as it organized the metadata from the various operational data definitions (ODD), assisted in the cleaning / scrubbing of the data via user-written code, and facilitated the creation of multi-dimensional databases (MDDB), datamarts and infomarts which were front-ended by exploitation tools including SAS/IntrNet™ based and SAS/EIS® derived applications. More importantly, the warehouse is under the central control of a single administrator and refresh is scheduled in a manner consistent with the needs of the department using the Windows AT scheduler (Windows NT 4.0). The current data warehouse will assist in multiple aspects of electronic transmission of regulatory data: data (SAS XPORT), reports, and computer assisted new drug applications (CANDA). The following SAS systems and products discussed included Base SAS®, SAS/GRAPH®, SAS/Warehouse Administrator™ v1.3, SAS/SHARE®, SAS/CONNECT®, SAS/MDDB™, SAS/IntrNet™ v1.2, and PH.Kinetics™ v2.00.

INTRODUCTION

The pharmaceutical industry, like other research and development (R&D) driven industries, generates large volumes of data of various types during the process of developing new medicines to ultimately improve the quality of people's lives. Data from early chemistry and pharmacology experiments, animal and human testing, and various processes involved with the manufacture of such chemical entities is collated into a regulatory document (NDA, new drug application). The NDA is reviewed by the Food and Drug Administration (FDA) and is the basis for approval of new medicines. As the information for a particular submission gets filtered and cleaned during the time the drug is being developed (11 years on average), this data is often extracted from other in-process data or historical data and prepared for regulatory submission in both detailed and highly summarized formats. More recently, the industry has embraced the value of maintaining a current, historical database of successes and failures with respect to potential drug candidates to facilitate knowledge-based decisions that call upon this ever-expanding R&D data environment. By the very nature of the typical infrastructure of a pharmaceutical company and the knowledge required to correctly assemble a database within a single data-generating department, these efforts have been disjoint and difficult to centralize. The advent of tools such as the SAS/Warehouse Administrator™ has improved the likelihood that such efforts can be linked and that risk-benefit analyses can be based on all pertinent data sources. DuPont Pharmaceuticals is at the genesis of efforts to link data from early discovery, preclinical and clinical development and eventually post marketing and competitive surveillance groups to improve the quality of decision making regarding new chemical entities and in-process drug candidates. The details of this "proof-of-concept" effort are contained herein using data from the Drug Metabolism and Pharmacokinetics Department's Clinical Pharmacokinetic Group. Data describing the safety and tolerability of new drugs in healthy volunteer populations is gathered during early phases of drug development. This data may also provide information on the pharmacokinetics (PK) and/or pharmacodynamics (PD) of a compound - generally defined as the effects of the body on the drug and the effect of the drug on the body respectively. Such data guides how new drugs are dosed in the patient population for which the drug is ultimately intended. Most recently, data demonstrating the differences in PK and PD due to age, Disease State, and other patient characteristics, which distinguish the patient population from the healthy volunteers, has been gathered and appreciated. The pharmaceutical industry in conjunction with the FDA has sought to examine PK and PD behavior in more relevant populations and in patients themselves when possible.

This desire has precipitated the need to pool and query PK and PD data across the numerous studies that may be conducted with a specific drug candidate. Population-based analysis is a model-based statistical approach used to examine the pharmacokinetic/pharmacodynamic behavior of new chemical entities in the patient population for which the drug is intended. It often encompasses study-related (SR) data (laboratory and clinical) from several trials of which the design may not be the same. Moreover, this technique involves the combination of data from several sources with predefined variable and file architecture based on protocol declaration. The creation of combined data set containing PK, PD, and SR data is a milestone in the timeline of the population analysis. The lifetime of such efforts typically occurs just prior to the NDA submission. Hence, the necessity of early planning and efficient coding is essential to: (1) provide "clean" data sets, (2) define partitions to construct randomized test and validation data sets, and (3) allow both construction of combined PK/PD/SR data sets and MDDBs for subsequent queries or incorporation into exploitation tools. The need to create population analysis data sets was the initial motivation behind the warehousing effort within the Clinical Pharmacokinetic Group at DuPont Pharmaceuticals. In actuality, it represents a single of many advantages provided by the combination of the PH.Kinetics™ and Warehouse Administrator™ products.

METHODS

Data Sources

Our goals were to build a data warehouse:

- To identify meta data for PK common repository
- To created detail tables, including SR, PD and adverse clinical effect (ACE) data
- To provide data for users to create ad hoc queries, analyses, and reports over the intranet
- Facilitate easier creation of population PK/PD/SR data sets
- Ultimately, to create data that can be ported directly to our electronic publishing systems (Core Dossier and Documentum), to be used in electronic submissions to the FDA and other regulatory agencies.

Data currently available in the warehouse site is derived from several departments including Drug Metabolism and Pharmacokinetics, Safety Assessment, Medicinal Chemistry, Clinical Pharmacology, and Biometrics. The Drug Metabolism and Pharmacokinetics department is composed of six different and diverse groups including Clinical PK, Pre-Clinical ADME/GLP, Discovery Support, Bioanalytical/Structural Chemistry, InVitro/Biotransformation, and Isotope Chemistry. Figure 1 illustrates the connection of the individual groups and their customers outside the department as well as the internal relationship of the groups.

DRUG METABOLISM & PHARMACOKINETICS DEPARTMENT

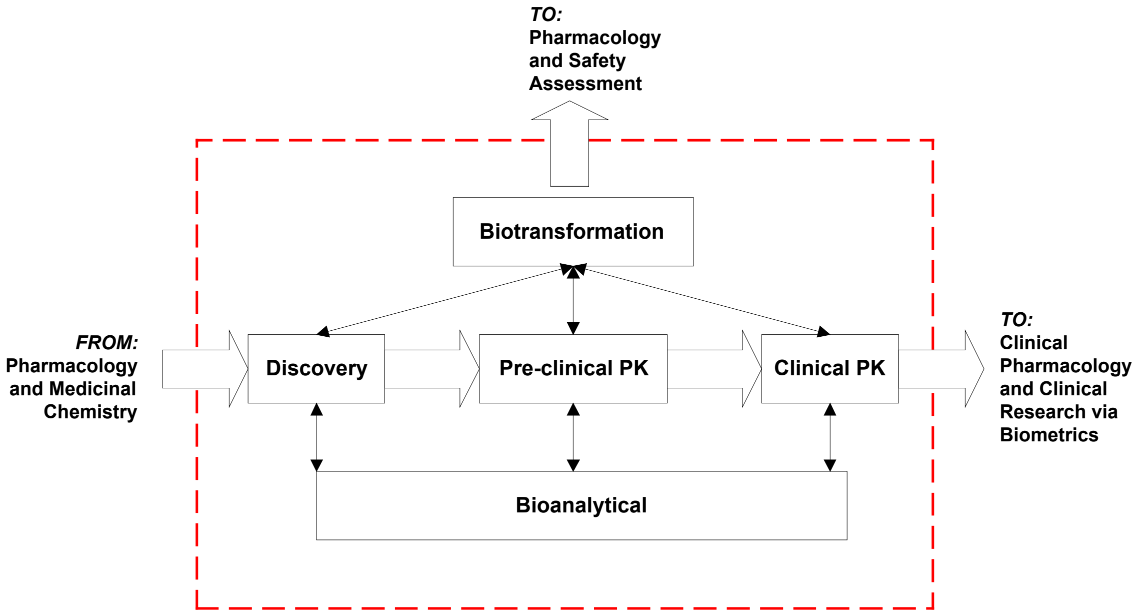

TO:
Pharmacology and Safety Assessment
Biotransformation
FROM:
Pharmacology and Medicinal Chemistry
Discovery
Pre-clinical PK
Clinical PK
TO:
Clinical Pharmacology and Clinical Research via Biometrics
Bioanalytical

Figure 1. DM&PK Department

Historically, data has been obtained from the clinical, pre-clinical, discovery and biometrics groups upon request in SAS files stored on the VAX Alpha running open VMS®, or as PC SAS data sets. We also get data from the same sources in the form of Microsoft® Excel worksheets and as Oracle® tables. In addition, we may get clinical and pre-clinical study protocol data directly from our Watson-Oracle® Laboratory Information Management System (Watson-LIMS) (See Figure 2).

The clinical and pre-clinical data typically include time, concentration and pharmacokinetic data. Pharmacodynamic, demographics, or adverse clinical events (ACE) data is obtained from the Biometrics Department. Demographic and adverse clinical event (ACE) data (SAS data sets) are merged across study. Pharmacokinetic data is either generated by our in-house, user written legacy Pharmacokinetic Information Management System, PK-IMS (a proprietary RS/1 application residing on a VAX cluster running VMS), CRO data, or through the use of PH.Kinetics™. The latter product is discussed in detail in a separate paper entitled, “The integration of PH.Kinetics™ with a clinical data warehouse.” Tables containing descriptions for a variety of data entries, such as randomization schedules, study medications and plasma / urine sample collection information is also stored as SAS files on VMS. Our target user platform is a Pentium PC (running Windows NT with a minimum of a 200 MHz processor and having a minimum of 64mb ram) connected to the company network via TCP/IP, which connects us to the VAX Alpha mainframe and other systems. Users will access exploitation tools through the company Intranet using SAS/IntrNet™ as well as the local area network (LAN). The company web-browser standard is Netscape Communicator® 4.6. The network operating system is MS-Windows NT Server, connected via SAS/CONNECT® and SAS/SHARE® software.

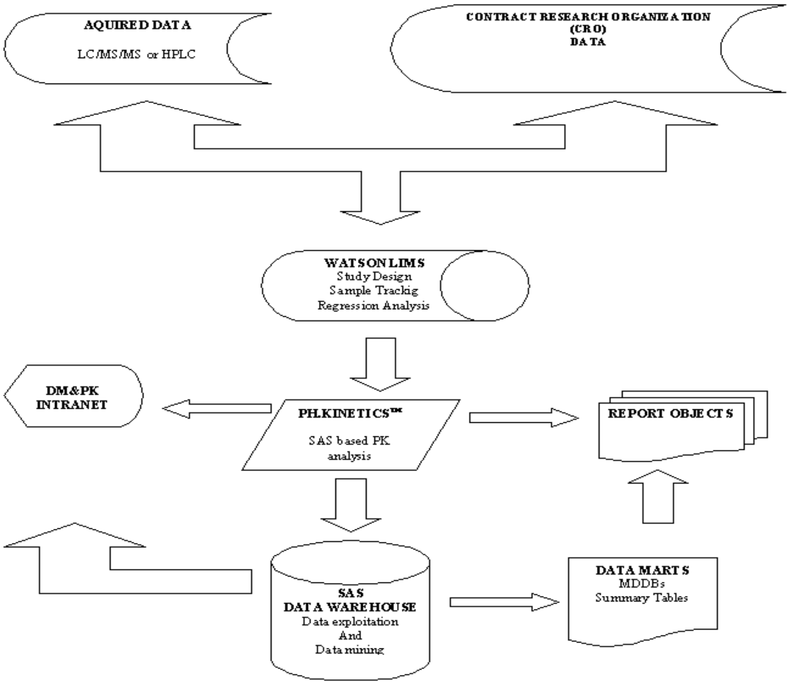


Figure 2. Data Sources

Transformation: Validation, Scrubbing, Integration, and Structuring

Transformation of PK and PD data involves data validation, data scrubbing, integration, and data structuring. Generating a frequency report is usually done to help with data validation or screening of the data to remove duplicate values, invalid, out-of-range, or missing values. Values obtained in the frequency report are then compared to values in a look-up table. Removing invalid, out-of-range, missing or duplicate values cleans data. Frequently, PK data has parameters defined in multiple attributes, such as, CL or CLO for clearance, AUCINF or AUCT for area under the curve from time 0 to infinity, or V or Vss for volume of distribution (see figure 3). Pharmacokinetic data parameters may have data values that are below quantifiable limits (BQL), BLQ, '.', or 0. Values listed as BLQ or BQL are recoded as 0, ensuring that the PK parameter remains as numeric input. Ensuring PK values remain as numeric input allows us to summarize PK parameters across compounds, studies or therapeutic areas.

Study	Gender	Race
DMP266_026	1	1
DMP266_027	2	White
DMP266_027	2	Black

Study	PK Parameter	Sex	Subject
266-026	CL	Female	1
266-027	AUCINF	Male	3
266-027	CLO	Male	2

Figure 3. Example of study, Gender, Race, and PK parameter inconsistencies that exist in PK data

Initially, study naming was not consistent across data sets. "Study" in demographics data set may have a data value defined as DMP266_026 while "study" in a concentration data set may be defined as 266-026. Another example of data cleaning occurs when we have to remove custom formats applied to a data set when the format is missing from the format catalog (see figure 4). Other cleaning tasks involve removal of unnecessary columns or the addition of columns to provide users with needed subject information. We removed columns such as CRF page 1, CRF comment 1. Age was added to the demographics data sets by defining AGE as "Date of Consent" minus "Date of Birth." Another feature of using the SAS/Warehouse Administrator™ software is the flexibility to insert user-written code at various stages of the cleaning process or simply to transform the data. This code is easily ported to run at many levels and hence code conservation is encouraged. Again, this facilitates the ability of a single administrator to maintain the warehouse.

```
libname userlib 'H:\clinical dw\user exit temp data files';
%let syslast=dt266.dt26629d;
data userlib.UE26629d;
     set dt266.dt26629d;
       format _doublee;
run;
%let syslast=userlib.ue26629d;
```

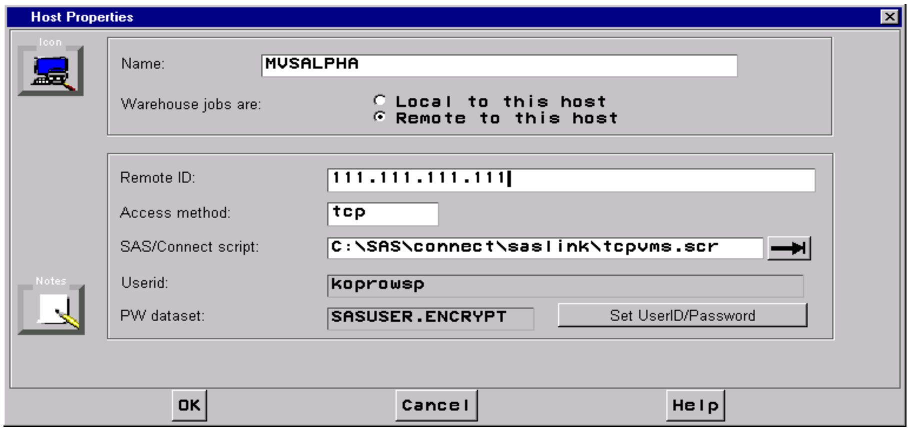
Figure 4. Example of user written code to remove a custom format

The Data Warehouse

There were five main steps used to construct the data warehouse for clinical pharmacokinetic data using SAS/Warehouse Administrator™:

- define a warehouse environment
- define Operational Data Definition (ODD) groupings and associated ODDs
- the warehouse elements-detailed and summarized data
- define the processes needed to move data into and through the warehouse
- load the warehouse

The physical location of the warehouse environment metadata is done with a LIBNAME statement. Shared metadata at the environment level, exists as library definitions for the location of warehouse elements, definitions of hosts (Local and VMS), contact information for the administrator and Investigators responsible for the data, and finally connections to the DBMSs (i.e., Oracle). Defining libraries for hosts or a DBMS is straightforward (see figure 5). Our SAS data libraries are collections of SAS files related by compound. SAS files come in several varieties. Most of the SAS files in our data libraries are data sets and data views. Usually, we create SAS data sets by importing data from other (external) sources such as Excel files (either Excel 97, Excel 95/5.0, or Excel for the Macintosh), Oracle® databases, or from other sources. The Excel files are created using the SAS import wizard, by using SAS/ACCESS® to PC file formats or using SAS/ACCESS® to ODBC.


Host Properties

Name: MVSALPHA

Warehouse jobs are: ○ Local to this host ● Remote to this host

Remote ID: 111.111.111.111

Access method: tcp

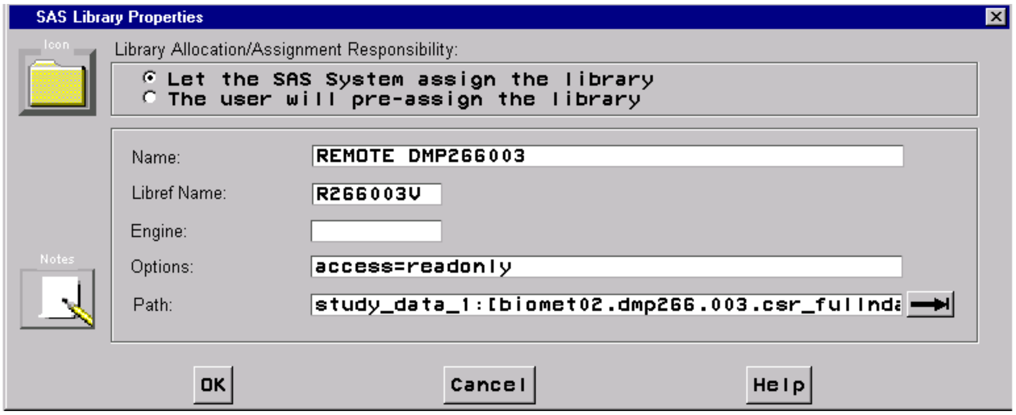
SAS/Connect script: C:\SAS\connect\saslink\tcpvms.scr

Userid: koprowsp

PW dataset: SASUSER.ENCRYPT

Figure 5. Example of defining a remote host with SAS/CONNECT

Some libraries are VAX directories (see figure 6) containing files and SAS data sets of related study information, or on the company file server in locked file directories. Each file is indexed with the study compound of interest.


SAS Library Properties

Library Allocation/Assignment Responsibility:

● Let the SAS System assign the library
○ The user will pre-assign the library

Name: REMOTE DMP266003

Libref Name: R266003V

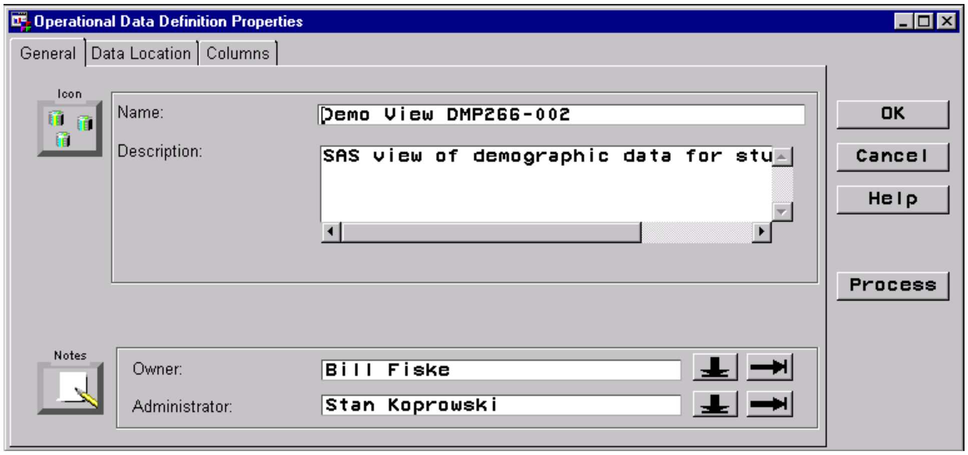
Engine:

Options: access=readonly

Path: study_data_1:[biomet02.dmp266.003.csr_fullnda

Figure 6. Example of defining a remote library

Our data warehouse environment shares three main operational data definition (ODD) groups; demographic, concentration, and pharmacokinetic. To define an ODD group to the environment you assign a name, description, an owner (Investigator) and an administrator (builder of the ODD). There can be an unlimited number of ODDs for each ODD grouping. Because the ODDs are stored at the environment level, they are available to any warehouse stored within the same environment. To define the SAS tables or SAS views that will store the data from the operational data source, you define the general, data location and columns tab of the ODD (see figure 7). The general tab will store information about the ODD's name, which is the owner, who is the administrator, and a description of the ODD. The warehouse elements or children may include data warehouses, subjects, summary, detail logical tables, and detail tables. A data warehouse can be thought of as repository of shared metadata among the warehouse elements. The warehouse holds large volumes of integrated read-only data. As a combination of SAS and non-SAS users, our data warehouse environment contains five warehouses, Anti-Inflammatory, Central-Nervous System, Cardiovascular, Virology and Radiopharmaceuticals. Within each warehouse, subjects, detail logical tables, detail tables, summary groups, summary tables, data mart groups, data marts, info marts, info mart items may be added. We will explore subjects, data mart groups, and data marts and summary items in further detail.



Name: Demo View DMP266-002

Description: SAS view of demographic data for stu

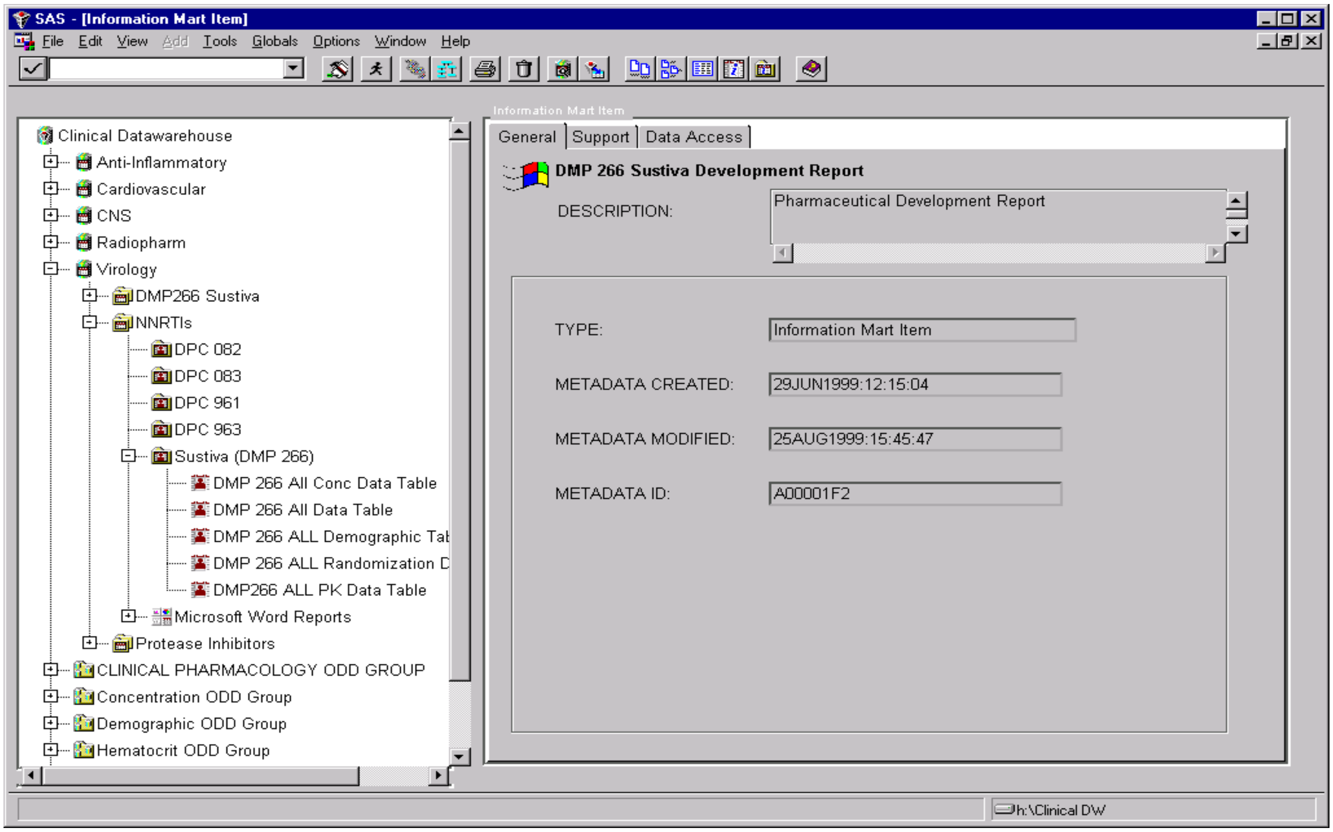
Owner: Bill Fiske

Administrator: Stan Koprowski

Figure 7. Example Operational Data Definition Properties Window

Therapeutic/franchise business areas (virology, cardiovascular, central-nervous system, anti-inflammatory, etc.) comprise the different warehouses. Subject groupings within a warehouse are by compound. A subject may be comprised of different types of data including any of the following: SAS data sets, SAS MDDBs, charts, reports, or graphs. Within the “General” tab, a subject name, description of the subject, an owner (Investigator in our warehouse), and an administrator define a subject. After all the subjects were added for a particular subject group, we then added Data Mart groupings (see figure 8). In our clinical pharmacokinetic warehouse, within a subject group, the Data Mart Groups are categorized by compound name. Categorizing the warehouse by therapeutic area and study will also facilitate future data mining needs.

Not wanting end users to access the actual physical data in the warehouse, we added data mart groups within each of the subject groups. The data mart groups were created within each subject grouping to hold summary data or tables containing the joined ODDs. Each data mart group was related to the subject. The compound name is a data mart group. One data mart was created for each for each ODD group. For example, data marts were created within the SUSTIVA™ data mart group for demographic, pharmacokinetic parameters, and concentration ODDs. Finally, a data mart was created that contained all other data marts, i.e.; a data mart (all SUSTIVA™ data) was created to contain the joined data mart items of demographics, PK, and concentration. After the data mart items were created, detail logical tables were created as grouping elements. To be consistent in naming, the names of the detail logical tables were the same as the data mart groups.



Clinical Datawarehouse
- Anti-Inflammatory
- Cardiovascular
- CNS
- Radiopharm
- Virology
  - DMP266 Sustiva
  - NNRTIs
    - DPC 082
    - DPC 083
    - DPC 961
    - DPC 963
    - Sustiva (DMP 266)
      - DMP 266 All Conc Data Table
      - DMP 266 All Data Table
      - DMP 266 ALL Demographic Tat
      - DMP 266 ALL Randomization D
      - DMP266 ALL PK Data Table
    - Microsoft Word Reports
  - Protease Inhibitors
- CLINICAL PHARMACOLOGY ODD GROUP
- Concentration ODD Group
- Demographic ODD Group
- Hematocrit ODD Group

DMP 266 Sustiva Development Report

DESCRIPTION: Pharmaceutical Development Report

TYPE: Information Mart Item

METADATA CREATED: 29JUN1999:12:15:04

METADATA MODIFIED: 25AUG1999:15:45:47

METADATA ID: A00001F2

Figure 8. View of Subject groupings across therapeutic area by compound data mart group

Detail tables were then added to the detail logical tables. A subject could only have one detail logical table assigned to it but may contain many detail tables. The major advantage of detail tables over data marts are that they can be defined once and linked to many different subjects. When defining the physical storage location of the detail table, SAS/Warehouse Administrator™, allows you to assign passwords for read, write, and alter access. You may also define indexes or add compression and encryption options. We did not compress our data due to the relatively small size of the data and the increase in CPU time required in processing the files. The encryption option allows you to impart security to the SAS data when outside of SAS software.

To this point in the construction process, individual warehouse elements (subjects, detail logical tables, data mart groups, etc.) and their associated properties have been defined but no relationships exist between the operational source data and tables within the warehouse and data have yet to be loaded into the warehouse. The final steps needed for constructing a data warehouse for clinical pharmacokinetic data are defining the relationships (processes) between the input source data and the target data, and loading the warehouse.
The four types of processes are:

- data transfers
- operational data mappings
- user exits
- record selectors

The most often used processes by our warehouse are user exits, data transfers, and operational data mappings. Unfortunately, for each warehouse process and load step in the flow, source code must be generated. Once input sources for the ODDs have been defined, data transfers from the remote host to the local host can be defined. A data transfer is required to move data from one host to another or when an input source and the target data reside on different hosts (see figure 9 for example of SAS/Warehouse Administrator™ generated source code).

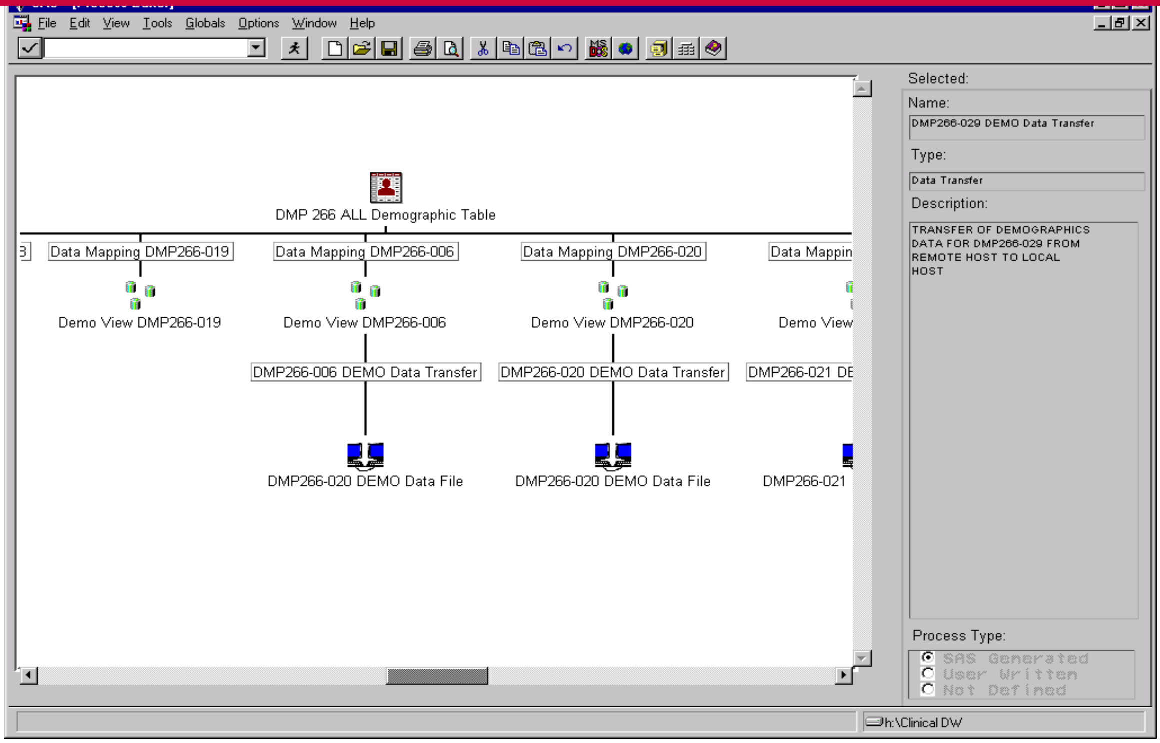
```
libname DT266 "H:\Clinical DW\Data Transfer";
options comamid=tcp;
%getusrpw(platform=MVSALPHA,macrovar=tcpSEC, data=SASUSER.ENCRYPT,destuser=koprowsp);
filename rlink "C:\SAS\connect\saslink\tcpvms.scr";
%let MVSALPHA=111.111.111.111;
signon MVSALPHA;
rsubmit MVSALPHA;
proc DOWNLOAD out=DT266.DT26629D status=no;
run;
endrsubmit;
```

Figure 9. Example of SAS/Warehouse Administrator™ generated code for data transfer

Operational data mappings sift multiple columns from one or more ODDs into a single detail table. Operational data mappings can be unions (ODDs that are appended, see figure 10), joins (ODDs are merged by one or more common columns) or simple merges (one ODD to a detail table).

SAS/Warehouse Administrator™ generates most of the code for loading the warehouse. However, these related but separate bits of code need to be linked together to form one set of code to enable the output of one process to become the input of the next process in the loading of the warehouse. SAS/Warehouse Administrator™ does this with use of a macro variable. The SYSLAST macro variable links the output data and input data for each process to be loaded into the warehouse. Typically, the source code is run in interactive mode, but the code can be saved (as a catalog entry or external file) and executed at a later time. This would be useful for refreshing the warehouse in a batch mode with a scheduler such as AT Scheduler.

Once the entire table has been loaded into the warehouse, summary groups, tables and MDDBs can be added. Summary data are used for querying and reporting, and therefore an important part of any warehouse. Summary data with class and analysis variables are easily generated with the software. Statistics can be added for the analysis variables. Available statistics include sum, count, average, min and max. In our clinical pharmacokinetic data warehouse we were easily able to view the minimum and maximum as well as, the average age of subjects in the clinical studies. We can also look at average pharmacokinetic parameters such as clearance, half-life, AUC, AUCtau, etc. across studies within a therapeutic class. Summary MDDBs are as easily created as summary tables. The MDDB could then be used in custom EIS applications, queried against with the output sent to a company Intranet for further exploitation.


File Edit View Tools Globals Options Window Help
DMP 266 ALL Demographic Table
Data Mapping DMP266-019
Data Mapping DMP266-006
Data Mapping DMP266-020
Data Mappin
Demo View DMP266-019
Demo View DMP266-006
Demo View DMP266-020
Demo View
DMP266-006 DEMO Data Transfer
DMP266-020 DEMO Data Transfer
DMP266-021 DE
DMP266-020 DEMO Data File
DMP266-020 DEMO Data File
DMP266-021

Selected:

Name:

DMP266-029 DEMO Data Transfer

Type:

Data Transfer

Description:

TRANSFER OF DEMOGRAPHICS
DATA FOR DMP266-029 FROM
REMOTE HOST TO LOCAL
HOST

Process Type:

● SAS Generated
○ User Written
○ Not Defined

h:\Clinical DW

Figure 10. Process Editor view showing example of Union or appending of multiple ODDs

RESULTS

DISCUSSION

The creation a clinical pharmacokinetic data warehouse within the Drug Metabolism and Pharmacokinetics section of the DuPont Pharmaceutical Company, has enabled us to have centralized control of the section’s data, consistency and unique generation of study tables and figures, elimination of redundancy, and increased analysis functionality through improved connectivity. We are now able to pool PK data with other data (demographic, safety, PD) across studies and across compounds within a therapeutic class. This data was then easily ported to EIS applications. We can perform further summarization and analyses on the data across the company Intranet (see paper entitled “Exploitation of Clinical Pharmacokinetic Data over a company Intranet: Strategies and Practices”). Eventually, we will be able to evaluate biopharmaceutical classification relative to PK/PD performance—in vitro/in vivo correlation. Ultimately, the creation of a clinical pharmacokinetic warehouse should enable us to make better decisions based on data and decrease development time.

REFERENCES

Building a Data Warehouse using SAS/Warehouse Administrator™ course notes, Chapters 1, 2, 3, 4, 5 and 6
Strategic Data Warehousing Principles Using SAS Software, Peter R. Welbrock, Chapter 4
SAS/IntrNet Software Version 1.2
SAS Software 6.12 TS055
SAS/Warehouse Administrator Software Version 1.3
Constructing a Data Warehouse, Curtis A. Smith Proceedings of the 24th SAS Users Group International

ACKNOWLEDGEMENTS

The authors would like to thank Tim Bara, Systems Analyst of the DuPont Pharmaceuticals Company for his help in configuring software, knowledge and expertise of VAX systems and network administration, as well as, his support and encouragement for this and other projects within the Drug Metabolism and Pharmacokinetics Department.

CONTACT INFORMATION

Stanley P. Koprowski, Jr.
DuPont Pharmaceuticals Company
Stine-Haskell Research Center
P.O. Box 30 (1090 Elkton Road)
Building 112
Newark, DE 19714

Questions: e-mail the author at Stanley.P.Koprowski@dupontpharma.com