Using the Data Warehouse Model to Streamline and Accelerate New Drug and Medical Device Development

Martin J. Rosenberg, Ph.D., MAJARO INFOSYSTEMS, INC.

ABSTRACT

Many companies that developed their own clinical data management systems in the 80's and early 90's settled on a model that enters and stores data in an On Line Transaction Processing (OLTP) oriented Relational Data Base Management System (RDBMS) and then moves that data to the SAS® System for analysis. This choice was made for a number of reasons: many information technology (IT) professionals did not consider SAS software to be a true relational database; companies selected their clinical data management systems at a time when the SAS system lacked certain relevant features found in traditional RDBMS systems; MIS departments involved in the selection of a database tended to chose the tools with which they were already familiar.

The original purpose of an RDBMS was to manage a company's operational data in the form of individual transactions. In recent years, IT professionals have come to recognize strategic information delivery as a second purpose for RDBMS systems and with this realization came a new term: the Data Warehouse. Although it still lacks certain OLTP features, the SAS system is an award winning Data Warehouse. With the advent of the Orlando II release of the SAS system (SAS 6.12), application developers now have all the tools necessary to construct clinical data management systems that match the very best systems built with OLTP software feature for feature, but which exceed them by facilitating the job of reporting the results of clinical trials to regulatory agencies. This paper contrasts the differences between OLTP and Data Warehouse systems and shows how using a Data Warehouse model and software such as the SAS system rather than an OLTP model can have important and beneficial consequences with respect to streamlining and accelerating new drug development, and integrating clinical data review and CANDA development into the clinical data management process.

DRUG DEVELOPMENT AND INFORMATION TECHNOLOGY

Unlike most other industries, companies in the pharmaceutical, biotechnology, medical device, and food products industries must first obtain government approval before they can bring a new product to the market. The process of obtaining government approval can often be long and complicated. For example, in the United States the research process for introducing a new drug can take five or more years and involves collecting animal toxicity data, information about the drug's safety and efficacy in humans, the stability of the drug (how long it can remain on the shelf without degrading); the pharmacokinetics of the drug, i.e., how it is metabolized in humans; and the ability of the company to manufacture the drug in production quantities. When all the required research is completed, the sponsoring company submits its information to the appropriate regulatory agency (in the United States the Food and Drug Administration or FDA) and makes a formal request for permission to market the drug or device in specific indications.

As can be imagined from the magnitude of information that must be collected, computer systems have long been a part of the development process. We call a computer system for the collection, retrieval, and analysis of clinical trials information, a Clinical Information System or CIS.

During the 1980's, as computers proliferated beyond Information Technology (IT) professionals to encompass most knowledge workers, the pharmaceutical industry and regulatory agencies such as the U.S. Food and Drug Administration sought to facilitate the drug development and approval processes through innovative uses of computer technology. One such initiative became known as a CANDA or Computer Assisted NDA. Similar initiatives were later launched in the U.S. for biologics and medical devices, and ultimately spread to other countries. Although an acronym for the U.S. drug approval process, in order to standardize on a single term for purposes of this paper, a CANDA is defined as any system designed to integrate modern information technology into the regulatory review process of new drugs, biologics, and medical devices.

While it is extremely difficult to measure the impact of CANDAs on the drug development process, the evidence that is available suggests that they can indeed accelerate the drug approval process as measured by elapsed time to a decision. For

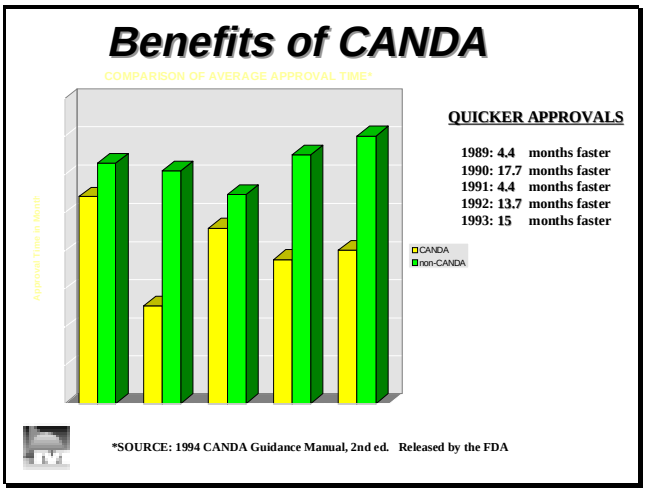

Figure 1: CANDAs reduce time to approval

example, in the *CANDA Guidance Manual* (FDA 1994) issued by FDA in 1994, a comparison of CANDAs that were submitted simultaneously with the respective paper NDA to CDER revealed average approval times that ranged from 4 to 18 months faster for the CANDAs (Figure 1).

Early CANDA efforts were expensive both in terms of the dollar cost to create the CANDA system and the extra time and manpower involved. Hence, companies are now looking for ways to incorporate the creation of a CANDA into the ongoing clinical data management process itself.

RELATIONAL DATABASE MANAGEMENT

The Relational Database model gained acceptance in the 1980s. In a relational database, data is stored in a series of rectangular arrays called tables. The columns of the tables are variables (or fields) and the rows are observations (or records) of those variables.

The original purpose of a relational database management system (RDBMS) was to manage a company's operational data in the form of individual transactions. When this is performed directly into the database without first completing a paper form, it is called on-line transaction processing (OLTP). Airline reservation systems and the NASDAQ stock exchanges are examples of OLTP systems.

Recently, a second purpose has emerged: the delivery of strategic information delivery through multi-dimensional analysis of operational data. This use of a database has come to be known as Data Warehousing.

An OLTP system is designed to perform the following:

1. High volume data entry. (The American Airlines Sabre reservations system performs more than one million transactions every 24 hour period).
2. Access individual records
3. Join data stored in different tables
4. Query records
5. Create reports and simple summaries

In contrast, a Data Warehouse is designed to perform the following:

1. Low to moderate volume data entry or access to existing machine readable files.
2. Access whole tables
3. Join data stored in different tables
4. On-line Analytical Processing (OLAP)
5. Multi-dimensional analysis

In comparing these two uses of databases, it seems clear that clinical data management, whose ultimate goal is to provide a clean database for statistical analysis, has more in common with Data Warehousing than with On-line Transaction Processing.

OLTP MODEL OF CLINICAL DATA MANAGEMENT

In the OLTP model of clinical data management, data capture is an "end" in itself. The goal is to rapidly enter and scrub data that has been collected on paper case report forms (CRFs). Upon completion, the data is transferred to the SAS system for statistical analysis and presentation in the form of reports and graphs.

This model complicates the processing of clinical data and creation of a registration package in several ways. First, Clinical Data Review (the use of CANDA like systems by in-house medical personnel) is not facilitated. The data resides in an OLTP system. Most of the review tools are only available with the SAS system. Hence, data must constantly be moved between the OLTP system and SAS.

Second, OLTP systems are designed to ensure that on-line transactions are not lost since there is no other record of the transaction. To accomplish this,

OLTP systems incur a substantial but necessary overhead each time data is accessed. When instead the data are being entered from paper and thereafter used in a read-only mode, the overhead is redundant and slows analysis.

Third, the use of OLTP systems creates a dual database situation. Data must be moved to SAS for analysis which takes up valuable time. Worse, since data is now stored in two locations, procedures must be put into place to ensure that both the official (OLTP) database and the analysis (SAS) database are kept synchronized as corrections are made. During the time crunch that often accompanies study analysis, this is not an easy task.

Lastly, CANDAs are not facilitated, and one has the task of deciding which copy of the database to send the regulatory agency.

CRITICAL PATH ANALYSIS OF CLINICAL DATA

Of all the steps involved in collecting, processing, analyzing, and presenting clinical data, patient enrollment and time on study is the longest. Compared to this, the time required to enter and clean the data collected is much shorter. For example, the majority of clinical trials collect 1,500 to 15,000 CRF pages over a 6 month to 2 year time frame. In contrast, a pair of experienced data entry operators can double key enter roughly 100 pages per day. This means a single pair of operators can enter all the pages in 15 to 150 days. For larger trials, data entry can be easily accelerated the low-tech way: by having more than 2 data entry operators.

Although the analysis plan and software programs can be developed in advance, statistical analysis cannot commence until the last data are collected, entered, and validated. Hence, statistical analysis is the rate limiting step on the critical path to filing. ***If you can eliminate the need to construct an analysis database, you can be ready to file sooner.***

This last point cannot be overstated. Suppose a drug is expected to sell 90 million dollars in the first year. For each day that approval is accelerated, first year sales increase by nearly 250,000 dollars.

If you could retain the desirable data entry, storage, and security features of an OLTP system, and add the analysis and reporting capabilities of a data warehouse, you could facilitate clinical data management and analysis, increase accuracy, and simplify development of a CANDA, all of which serve to accelerate filings. Hence, it makes sense to explore the use of Data Warehouse systems to perform clinical data management.

THE SAS DATA WAREHOUSE

As currently practiced, the ultimate goal of clinical data management is to transcribe, with as much accuracy as possible, data collected on paper CRFs. The CRFs are legal “source documents” whose accuracy and veracity is attested by the investigator. Even in those situations where data is entered directly into a computer, such as remote data entry, paper CRFs are generated and signed by the investigator. Hence, clinical data management is not OLTP. While the SAS system still lacks certain features necessary for high volume OLTP, the Orlando II Release (6.12) has all the features necessary to allow developers to create clinical information systems that match the data entry capabilities of the best OLTP systems feature for feature.

Due to its powerful tools, the SAS system is already part of the clinical data management and analysis process. Many companies have considerable SAS expertise resident in their shops. There is increasing recognition of SAS as a data warehouse in the IT industry. For example, the Orlando (6.11) release was voted as Data Warehouse Product of the Year for 1996 by Datamation magazine. Hence, it makes sense to leverage existing resources and choose the SAS data warehouse to construct a clinical data management system.

INTEGRATING CANDA DEVELOPMENT INTO THE CLINICAL RESEARCH PROCESS

CANDAs have evolved into three components: data, image, and text. The data component is comprised of the database, analysis programs, and tools to let medical and biometrics reviewers access the database. The image component consists of images of the CRFs and possibly the entire registration package. The text component often means a copy of the registration package stored in word processing files so as to permit cutting and pasting from the NDA into the documents that need to be prepared by the reviewer. With careful planning and the use of the Data Warehouse model, it is possible to integrate the data and image components into the

clinical data management operations already being performed.

As CRFs are collected, it is common to stamp them with a unique document ID number (also known as an accession number) and log them into the database as received. Increasingly, companies are also scanning the CRFs into an image database (Figure 2). Image databases have several advantages: the original CRF can be safely stored; many people can access the CRF at the same time, e.g. data entry and monitors; and it provides a single source for the most up to date copy of the CRF.

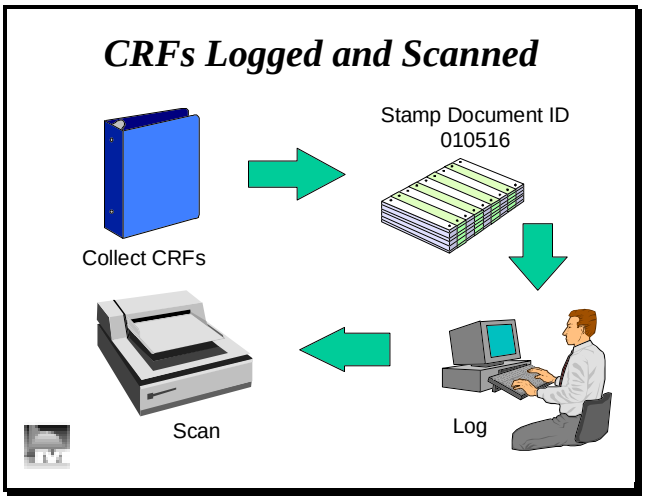

Figure 2: Logging and Scanning of CRFs

The log-in process, used in clinical data management to track CRFs, can also serve as the index required by the image database to display a CRF image and data entry screen side by side in order to facilitate data entry. As data entry is performed, the links between the records in the database and the CRF images are created. Hence, the database and CRF image components of the CANDA can be created by substantially leveraging the effort already being expended in clinical data management.

CLINACCESS/POWERSERVER™

The data warehouse clinical data management model has been implemented in ClinAccess/PowerServer™ software, an integrated clinical trials system that combines the data entry and data management capabilities of traditional clinical information systems, with clinical data review features that permit monitors, CRA's, managers, and other members of the clinical staff to monitor the progress and quality of ongoing clinical trials.

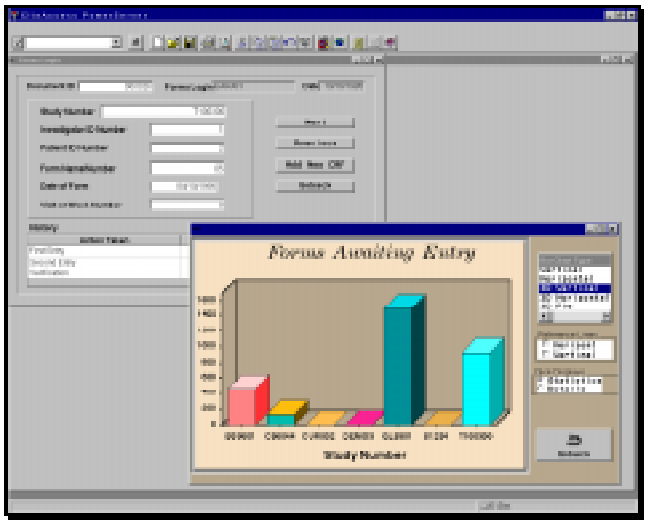

Figure 3: Forms Tracking system logs CRFs and displays up to the minute status

Advanced features include a forms tracking system that logs each CRF page and tracks its status throughout data entry and validation. A graphical display shows up to the minute status and offers drill-down capabilities for details about specific studies (Figure 3).

CRF pages can be scanned, linked to the database, and displayed side by side with a database screen for data entry or review (Figure 4).

The Query Management system allows operators to enter data clarification requests which are linked with queries generated by batch validation reports and tracked through resolution (Figure 5).

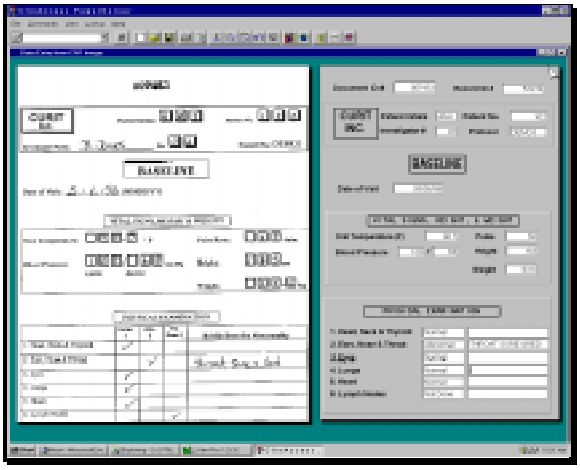

Figure 4: Displaying CRF image and data base screen side by side

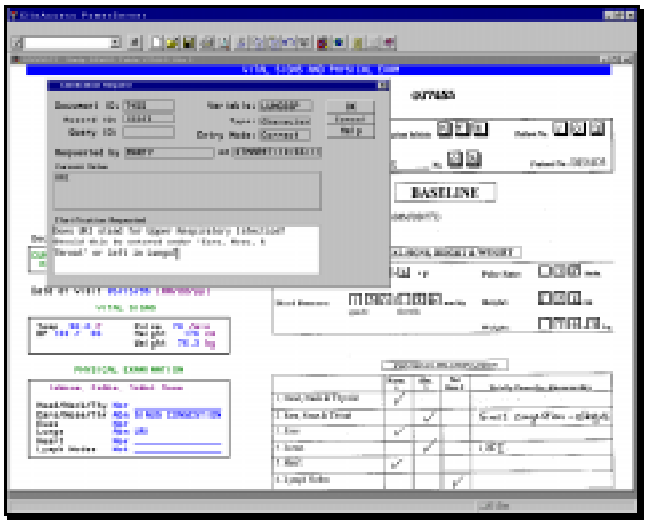

Figure 5: Data Clarification window

SUMMARY

The CANDA initiative begun in 1986 has advanced to the stage where nearly 1 out of 3 new submissions have a CANDA component. In order for CANDAs to continue to become a routine requirement, the creation of CANDAs must be simplified and begun early in the clinical development process. The use of a Data Warehouse model instead of an On-line transaction processing (OLTP) model is an important step in this direction.

ClinAccess/PowerServer™ software is the first major clinical trials system to implement the Data Warehouse clinical data management model. Developed entirely with SAS software, ClinAccess is designed to provide monitors, CRAs, and other non-traditional users with access to the information stored in clinical databases. ClinAccess provides: single or double-key data entry; a data dictionary, audit trails, forms tracking, query management, image processing, viewing and querying of data; graphics; descriptive statistics; and report generation.

REFERENCES

Food and Drug Administration (1994). *CANDA Guidance Manual, Second Edition*. U.S. Department of Health and Human Services, Public Health Service, pp. 50.

Rosenberg, Martin J. (1996). ClinAccess™: An Integrated Client/Server Approach to Clinical Data Management and Regulatory Approval *Proceedings of the Twenty-First Annual SAS Users Group International Conference*. SAS Institute Inc., Cary, NC. pp. 1190-1198.

ACKNOWLEDGMENTS

MAJARO INFOSYSTEMS, INC. provides statistical and information management services to the pharmaceutical, biotechnology, medical device, and food products industries, and specializes in extending computer technology to non-traditional users.

For further information regarding this paper, please contact:

Martin J. Rosenberg, Ph.D.
MAJARO INFOSYSTEMS, INC.
2700 Augustine Drive
Suite 230
Santa Clara, CA 95054

tel. (408) 562-1890
fax. (408) 562-1899