Paper DM08

Managing Clinical Trials Data using SAS® Software

Martin J. Rosenberg, Ph.D., MAJARO InfoSystems, Inc.

ABSTRACT

For over five years, one of the largest clinical trials ever conducted (over 670,000 pages from 90,000 patients and 2,000 investigators) was managed using a SAS® Powered clinical data management system. All data were entered, stored, managed, and analyzed using SAS/SHARE® as the database server, SAS tables for physical storage, and Microsoft Windows® as the operating system on both the server and desktop. Electronic images of Case Report Forms (CRFs) were used extensively in lieu of the paper CRFs.

Although it origins are in statistical analysis, the SAS system has evolved to encompass many features once thought to be the exclusive province of traditional transaction oriented relational database management systems. These new capabilities include dictionary level integrity constraints, declarative referential integrity, built-in audit trails, concurrent data entry via a database server that provides true row-level locking, SQL, virtual tables (views), indexing, compression, encryption, and interfaces to industry standard application development tools such as .NET and JAVA.

This paper discusses the features of SAS that make it a powerful environment for clinical data management, our experience in managing large trials using SAS, and our vision for increased roles for SAS as CDISC and electronic submissions become the norm.

INTRODUCTION

An urban legend has developed that the SAS system is a good tool for small studies, but that a commercial online transaction processing (OLTP) oriented relational database is required to manage large amounts of data.

For over five years, MAJARO was engaged by Mentor Corporation (Santa Barbara, California, USA) to manage data from the ADJUNCT Study of Silicone Gel-Filled Breast Implants. This study evaluated the safety of the implants in women undergoing breast reconstruction or revision of existing implants. The ADJUNCT Study was one of the largest clinical trials ever conducted: more than 670,000 pages were collected from 90,000 patients and 2,000 investigators. All data were entered, stored, and managed using ClinAccess/PowerServer™, a SAS Powered 21 CFR Part 11 compliant clinical data management system that utilizes SAS/SHARE as the database server and SAS tables for physical storage. Due to the size of the study, a number of innovative approaches were employed, including extensive use of casebook imaging.

In April 2005, an advisory committee to the U.S. Food and Drug Administration (FDA) recommended approval of the implants. In July 2005, FDA issued an “approvable letter” subject to conditions that are generally consistent with the advisory panel's deliberations.

CONSIDERATIONS FOR MANAGING LARGE STUDIES

Due to the size of the study, efficiency was an importation consideration. A team of 25 data entry operators would be entering the data on a daily basis, with concurrent access additionally required by data managers, supervisors, and programmers. The handling of 150,000 pages of paper per year would alone be a daunting task. To overcome these obstacles, a number of innovative techniques were successfully employed.

Casebook Imaging

Rather than enter data from paper Case Report Forms (CRFs), CRF images were used for all data management activities. Each CRF page had a unique pre-printed document ID number. Upon receipt, the CRFs were scanned by the sponsor into TIFF files and then stored. TIFF files are an industry standard for storing graphical images and can be displayed by all versions of Microsoft Windows. The TIFF files were copied onto CDs for transfer to MAJARO. Up to 30,000 images can fit on a single CD, although we found it convenient to receive 2,000 – 3,000 images at a time. All subsequent processing was paperless and used only the images to manage the data.

Due to the constant stream of large numbers of casebook images, the first task was to record the receipt of each page in a database to avoid loss. The image files were copied from the CDs onto MAJARO's network. Each CD was given a unique identifier and processed as a batch. Once loaded onto the network, the images were indexed using ClinAccess/CaseBook™, an imaging product that integrates with the ClinAccess/PowerServer™ database. This involved recording identifying information such as the batch number, document ID number, study, patient ID, investigator, form name, and visit date.

Casebook imaging had a number of advantages:

- Paper need be handled only once with reduced risk of loss or damage to the pages.
- Images can be viewed by any number of people simultaneously, and if required, from remote locations.
- Images can be retrieved in a random order. In particular, CRFs could be received and scanned sorted by patient, but entered by form without the need to sort or collate the pages. As borne out by our experience, data entry operators can enter data many times faster when entering a single type of form, than when switching between forms after each page.
- Images could be annotated electronically and be either printed, faxed, or emailed to the investigator for resolution of queries.
- Up to 99 revisions of each page could be handled. The system can be configured so that the most recent revision is always displayed first.

Forms Control

The ClinAccess/CaseBook™ index is imported into the ClinAccess/PowerServer™ Forms Control system which tracks each page throughout the data management process. As pages are entered, the following information is automatically captured in Forms Control: data management action (e.g. first entry), user, date, time, the name of the database table, and unique identifiers of the rows entered from the page.

With 670,000 pages, it would be impractical to track pages manually. However, with the Forms Control database, we could easily track the status of each CRF page in real time, and create computer generated reports to ensure that each page completed data entry.

For convenience, the Forms Control window links information about each page with the image of that page (Figure 1). Lookup tools make it easy to display specific pages for each patient.

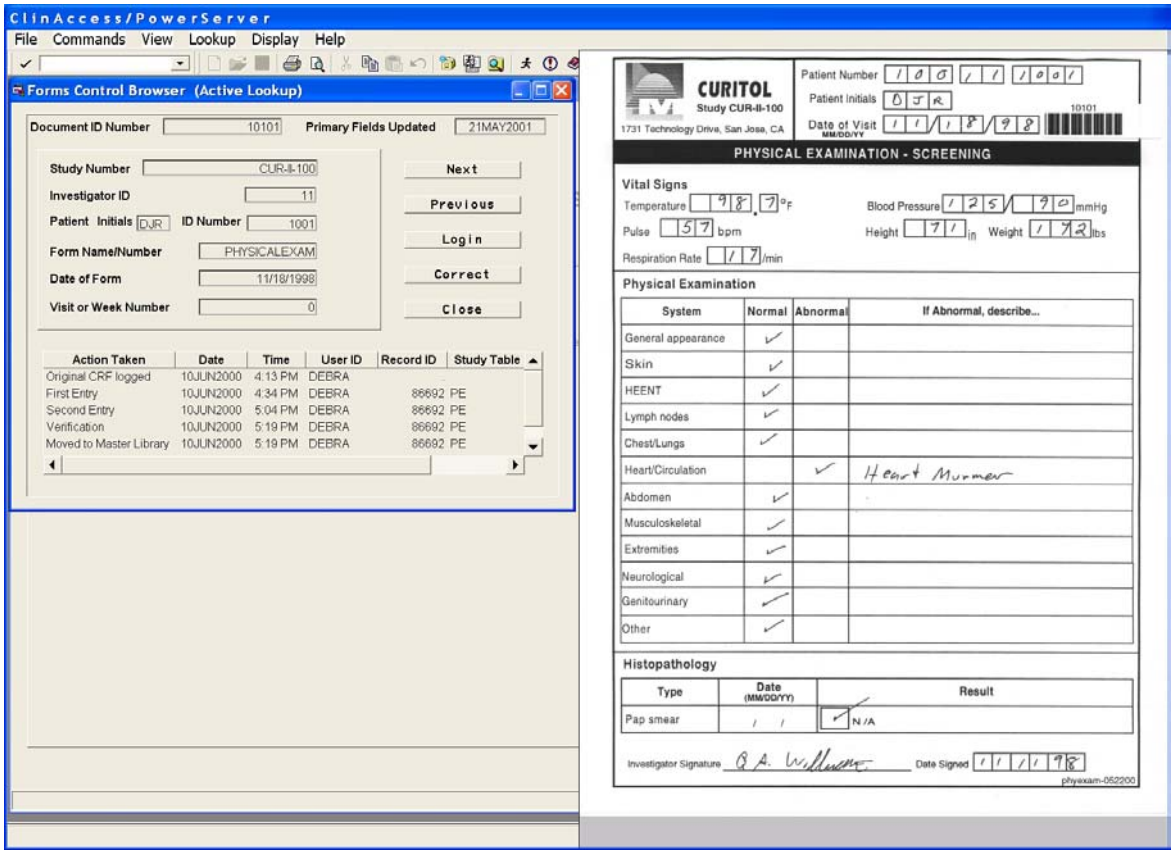

Figure 1: Forms Control window displays progress of each CRF page

Data Entry

Data were double-key entered using split-screen data entry that displayed the image side by side with the data entry screen (Figure 2). The image to be entered is brought up automatically and linked permanently to the study data. In addition to the obvious convenience of paperless data entry, imaging allows the data entry operator to enlarge sections of the page and to bring up any other page for the patient.

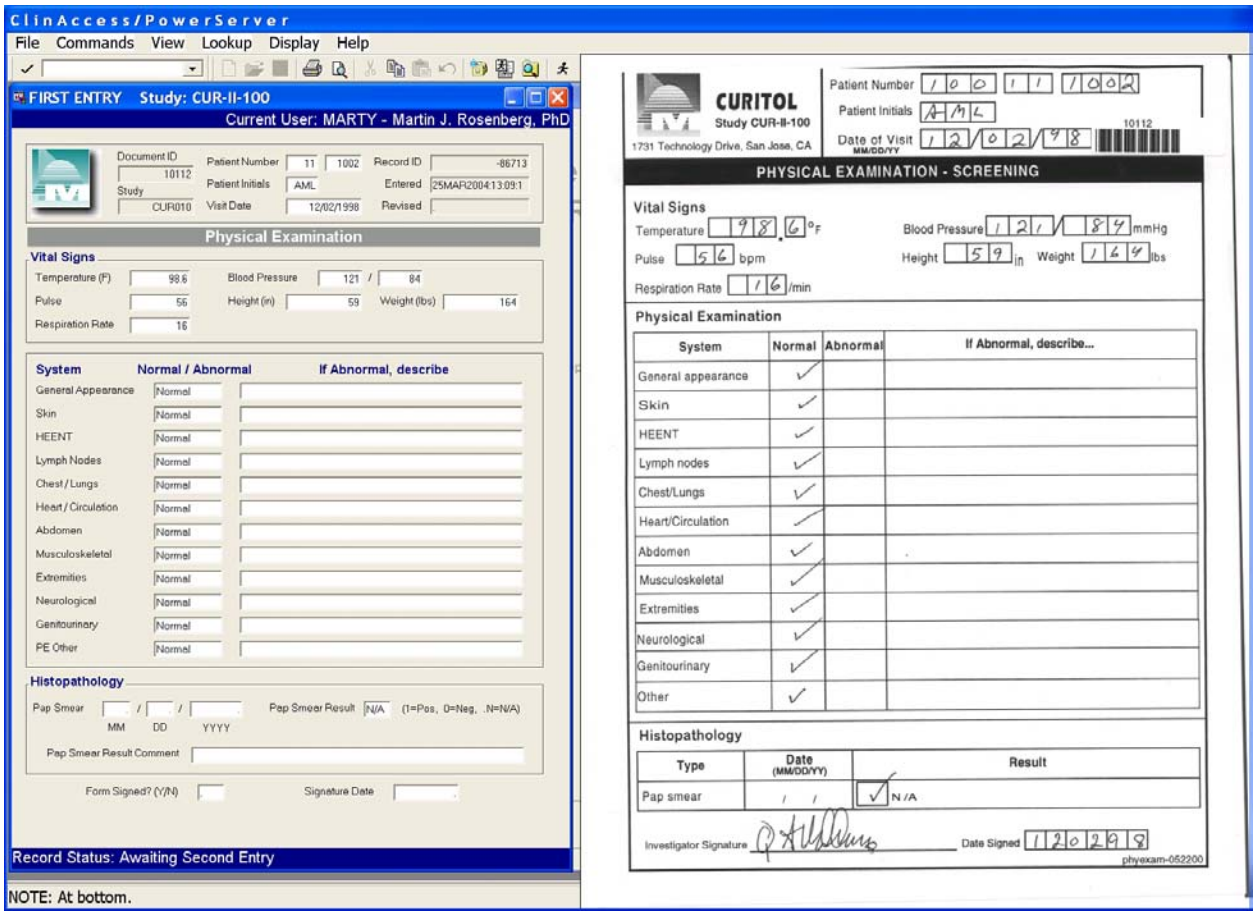

Figure 2: Paperless split-screen data entry

SAS/SHARE Database Server

A database server is a program that negotiates requests from multiple users to access and update data stored in a database. SAS programmers know that any number of users can simultaneously obtain read-only access to data stored in SAS tables. SAS/SHARE extends this ability to allow any number of users to have write or read access to SAS tables. Technically SAS/SHARE is a SAS session that runs continuously on a server. When a user requires update access to an existing row of the table in order to make changes, the SAS/SHARE server "locks" the row so that only that user can modify the locked row. Other users can lock other rows of the table, and all users have read access to all rows of the table: both the locked and unlocked rows. So for example, a programmer could be running reports at the same time that other users performed data entry.

Hardware and Operating System

We used a common Windows – Intel architecture for the server and network operating system. When selecting a computer to function as a database server, there are several considerations. While processor speed is important for any computer, hard drive speed and RAM can have a greater impact on computers used as database servers.

1. The speed of a hard drive can be measured two ways. Access time measures the average time it takes to position the read head on the desired sector of the drive. For database use, the more important measurement is rotational speed measured in revolutions per minute (RPM). Pick drives that spin at least 10,000 RPM, preferably 15,000 RPM. (In contrast, the drives used in most desktop computers spin at 5,400 or 7,200 RPM).
2. Use a RAID 5 configuration for the hard drives. This both increases speed and adds redundancy, so that if one drive fails, the server continues to function without loss of data. When the drives are "hot swappable" a failed drive can be replaced without the need for shutting down the server.
3. Use as much RAM as possible. 32-bit Windows operating systems can address a maximum of four gigabytes (4 GB) of RAM. (This limit is increased more than a 1000-fold by the new 64-bit Windows operating systems.)

Presently, we use a department quality server with a 3.0 GHz Intel Xeon processor, 4 GB RAM, and a series of SCSI hard drives in a RAID 5 configuration, and the Windows 2003 Server operating system. This system was more than adequate to service 25 people entering data into the same table simultaneously with no degradation of performance. Such a computer can be purchased for five to ten thousand U.S. dollars, depending on the amount of the hard-disk storage and level of built-in redundancy.

SAS PROGRAMMING TECHNIQUES FOR CLINICAL DATA MANAGEMENT

When using SAS for data management, one can increase performance using techniques not ordinarily employed during analysis. In particular, two techniques allowed us to achieve extremely rapid retrieval of data.

Indexing

Database indexes function much like the index of a book. When a variable is indexed, SAS stores additional information concerning the location of each value of the variable. When searching for rows that have a particular value of the indexed variable, SAS first looks in the index and then can go directly to desired rows of the database table. Without indexing, SAS would need to search the table sequentially from the beginning until it locates the rows that satisfy the search criteria. SAS indexes are defined using the DATA step, PROC SQL, PROC DATASETS, or the SCL ICREATE function

WHERE Clauses

When SAS/SHARE receives a request for data from a SQL query, DATA step, or PROC step that uses a WHERE clause, the SAS/SHARE server evaluates the query and returns only the data that satisfies the WHERE clause. Since data moves much quicker within a computer than across a network, retrieval speed can be considerably enhanced over other techniques such as the Subsetting IF statement of the DATA step. When a WHERE clause is used in conjunction with indexed variables, the increase in response time can be even higher.

SAS RELATIONAL DATABASE FEATURES

Although it origins are in statistical analysis, the SAS system has evolved to encompass many features once thought to be the exclusive province of traditional transaction oriented relational database management systems. These new capabilities include:

- Dictionary level integrity constraints – validation rules built into the definition of the SAS table at the time the table is created
- Declarative referential integrity which creates links and defines relationships between database tables that are part of the same database
- Built-in audit trails. Although the built-in audit trails do not satisfy FDA requirements to capture the reason for the change on a per field basis, they are nonetheless useful for recovery purposes and for satisfying the 21 CFR Part 11 requirement that data should not be able to be changed without leaving a trace in an audit trail
- Concurrent data entry with true row-level locking
- SQL, the industry standard language for database retrieval.
- Indexing for rapid retrieval of data
- Compression to save storage space
- Encryption to eliminate the ability to circumvent SAS security by using operating system utilities
- Interfaces to industry standard application development tools such as .NET and JAVA via SAS Integration Technologies (IT).

ADVANTAGES TO USING SAS AS A CLINICAL DATA REPOSITORY

The typical purpose of most corporate databases is to store and retrieve individual transactions. Banks need a record of each deposit and withdrawal. Airlines need a record of each seat on a flight, to determine whether any seats remain available for sale. Stock exchanges need to know who sold a stock and who purchased it. Credit card companies need to know each sale charged to each credit card. All companies need records of income and expenses.

In contrast, the primary purpose of the clinical trials conducted by pharmaceutical and medical device companies is to perform sophisticated statistical analyses that support product approval. While we collect information from individual patients, storing the information is not an end in and of itself. Instead, our primary interest is to understand the experience of the treatment group as a whole. The time savings of eliminating the need to extract, transform, and load data from an OLTP database into SAS can be considerable. Hence, the most obvious advantage to using SAS as a repository is that the data are ready to be used for analysis. However, there are many other advantages.

In online transaction processing (OLTP), it is critical that either the entire transaction be recorded or no part of the transaction be recorded. For example, the sale of stock on a stock exchange requires that both the sales portion and the purchase portion of the transaction be recorded. While commercial databases contain tools to support this requirement, these tools and procedures do add to the complexity of creating and maintaining databases.

Clinical data management does not require all or nothing transactions. Hence, SAS is an excellent fit for this application, and systems such as ClinAccess™ that are built on the SAS platform can enjoy the following benefits:

- Lower ongoing costs by eliminating the need for a full-time database administrator
- Rapid creation of new study databases
- Lower requirements for computer resources
- Enhanced tools for cleaning data
- Enhanced reporting tools

IMPACT OF CDISC AND ELECTRONIC SUBMISSIONS

With the advent of the CDISC standards and a regulatory climate that permits and even encourages paperless submissions, the inclusion of the study database as a permanent part of the regulatory submission, is going to become commonplace. FDA has announced plans to construct a database known as JANUS that spans submissions across sponsors. The Biometrics branch of FDA looks to the CDISC ADaM model for database tables that are "one-PROC away" from analysis. Although not often discussed, companies that maintain two databases – the "official" copy in a commercial database and the analysis copy in SAS – may increasing need to spend extra time and money to ensure the data is identical in both places. In practice, keeping two copies of data identical when they are stored in incompatible databases can be very difficult. Using SAS for data capture and storage, in addition to analysis, can eliminate this extra effort.

CONCLUSION

SAS has long been the de facto standard for statistical analysis of clinical trials data. While many companies have used SAS to manage clinical trials data, there remains a perception that SAS cannot handle large quantities of data. MAJARO's experience using ClinAccess™, a SAS Powered clinical data management system, shows that such systems can scale to handle even the largest studies conducted by industry, while simultaneously reducing time to study setup, time from database lock to analysis, and ongoing maintenance costs.

REFERENCES

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

CONTACT INFORMATION

MAJARO InfoSystems, Inc. provides software and CRO services to the pharmaceutical, biotechnology, and medical device industries and develops ClinAccess™, the leading SAS Powered clinical data management system.

For further information regarding this paper, please contact:

Martin J. Rosenberg, Ph.D.
MAJARO InfoSystems, Inc.
2350 Mission College Boulevard
Suite 700
Santa Clara, CA 95054-1552
USA

+1 408-330-9400 phone
+1 408-330-9410 fax
www.majaro.com web site

Address email to mrosenberg at the same domain name as the web site.