

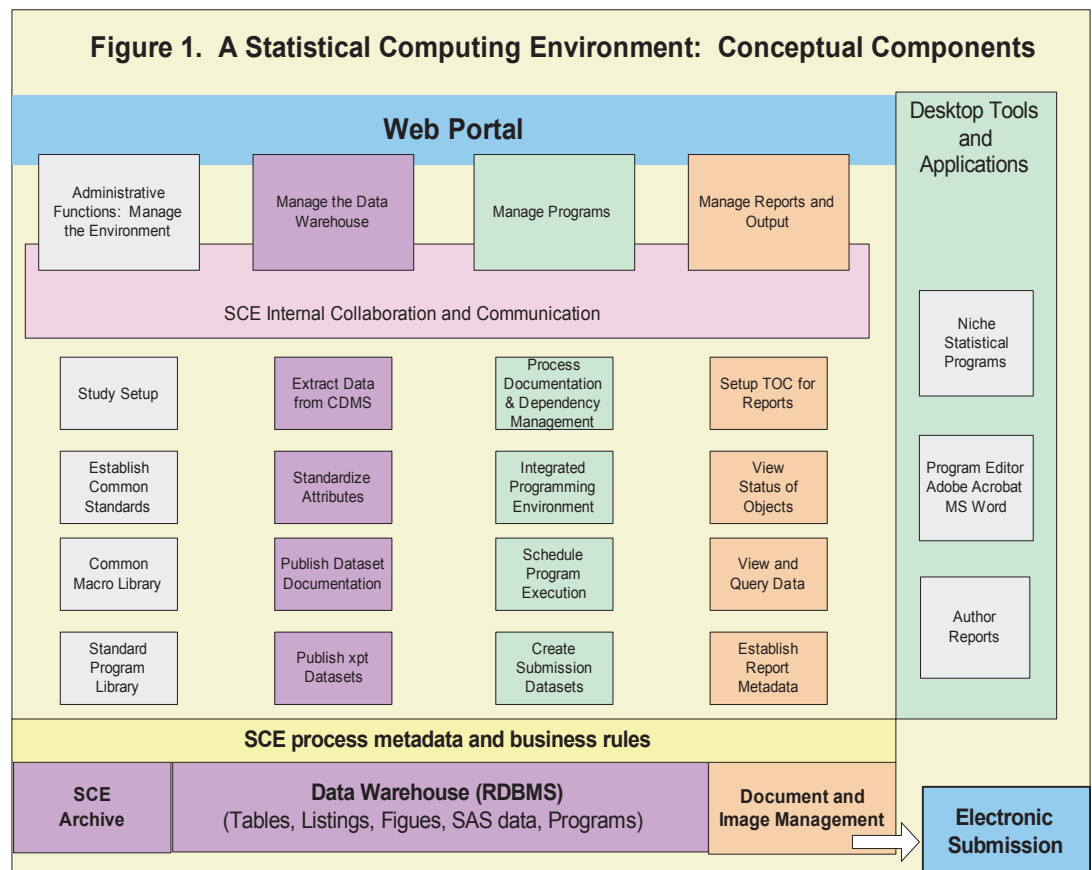
The Statistical Computing Environment

Considerations for building the infrastructure to assure quality electronic submissions and meet regulatory requirements

Statistical analysis activities conducted by clinical biostatisticians and clinical data programmers are required to interpret clinical trial data. Unlike clinical data management systems designed for squeaky clean accountability in all phases of the data entry and management functions, clinical biostatistics processes are largely *ad hoc* and tedious. Integrated systems are needed by industry to provide a programming environment for production of tables, figures, and listings, for storage of data in secure and compliant repositories, for easy data access to data and results in a collaborative environment, and electronic submission of results to regulatory authorities. The whole data analytic process needs to be transparent and traceable. Here we will discuss some of the issues related to specification, execution, quality, and documentation of statistical computing for clinical trials.

We refer to the statistical computing environment (SCE) as a flexible system for creation, control and documentation of the elements specified in the statistical analysis plan for a clinical trial. A collection

Figure 1. A Statistical Computing Environment: Conceptual Components



Key Components:

- The Analysis Plan
- Analysis Data Files
- The Programming Process: statistical results
- Report compilation
- Compliance and Validation; results verification
- Publish results and data
- Data viewer

The Statistical Plan:

- Detailed statistical methods
- Analysis file specification
- Supportive tables, figures, and listings

SCE Concepts:

- Metadata foundation
- Data transformation and standardization
- Standard reporting software
- Manage Components:
 - Analysis files
 - Programs
 - Reports
- Process documentation
- Publish data
- Audit trail
- File security
- Version control
- Study archival

21 CFR Part 11 Compliance

of tools and operating procedures comprises a statistical environment for computing in a regulated environment. The output of the environment are controlled documents published in portable document format (PDF). The statistical tools used are primarily SAS Software and S-Plus for tables and graphics.

The conceptual components of the environment is depicted in Figure 1. A key to developing a structure is the administrative function which allows a study setup which identifies the objects to be produced and tracked. Underlying the administrative structure is process metadata available to the programs. Using the metadata program status such as validation milestones can be tracked. An additional aspect of the process which departs from standard practice is a data repository which is based on a relational database management system so track versions of the SCE objects and to enforce security rules.

Implicit in Figure 1 is the ability to easily query data in the data repository and view results via a web browser. This is the easy way to make the statistical results accessible to other team members. The ability of team members to view raw data directly and query subsets can cut down on special requests for data.

Statistical Analysis Files

Data is typically collected in a logical forms-based format and stored in a relational database. These data are usually stored in a highly normalized form not consistent with the ultimate use of the data. Therefore the operational data needs to be transformed to analysis data sets that are easily used for statistical analysis. These files include derived variables. Leverage standards such as the Consortium for Data Interchange Standards (CDISC) ADaM model for preparing analysis files and exporting data to the FDA.

Usually multiple analyses data sets are created. An individual analysis file may be in one of several formats depending on the type of data and the planned type of statistical analysis. Ideally the analysis files can be used as input directly to statistical procedures. The data from the operational database must be transformed into one of these formats.

Creation of the analysis files is a critical step in the reporting process that presents the greatest opportunity for error. Merging files of different shapes and creating derived analysis variables are prone to error. Creation of analysis variables may be conceptually easy, but the data often wreak havoc with coding due missing data and unexpected data abnormalities. For this reason, PharmaStat recommends that a data integrity report be issued by the data management department for use by programmers of statistical tables to help anticipate special programming driven by the data. The complexity of this process and its impact on conclusions make a compelling business case for program validation.

The Programming Process: Statistical Results

The statistical results usually take the form of tables, figures and domain data listings. Several objects need to be managed here: the programs for data manipulation and statistical analysis, the analysis data sets, the results of the analysis, and the integration of the statistical results into a publishable report. How do we manage all of these objects in a way that will guarantee the traceability and transparency we must have from this process? The best way is to work in an environment that tracks all of the objects. By developing a table of contents of the objects to be created one can track the objects. The table of contents itself becomes a part of the study metadata.

The environment should also include facilities for program code management, dependency and object management, etc. The environment would typically include standard programs and algorithms for producing common reports of trial data. Above all, the statistical computing environment develops electronic documentation of the entire process. The SCE creates a workflow and associated job control for activities related to the statistical analysis.

Validation and Compliance

PharmaStat believes that most pharmaceutical and biotechnology companies are out of compliance with the requirements of CFR Title 21 Part 11 the electronic records and signature rule in the area

of clinical trial statistical data processing. In addition, many lack a systematic quality approach to this important area of clinical trials.

To comply with the regulations, the environment should be secure, allowing only authenticated individuals access. There must be version control and audit trails on the contents of the data repository and the ability to produce reports based on the audit trail and changes. Electronic signatures need to be implemented and used for critical documents. For these reasons, we believe that all objects should be stored in a secure relational database management system.

Systems Characteristics of a SCE Architecture

A typical architecture is displayed in Figure 2 below. The SCE client is a Windows machine. The SQL database server is analogous to an ordinary file server except that the data, programs and output will be stored on the SQL database server with security, file versions, and audit trails. This helps to meet the Part 11 requirements. Validated compute servers are used to run statistical analyses. Non-technical users may access results and perform data queries through a web interface. The environment may also have links to clinical database systems for uploading data for analysis or document management systems for publishing depending on the corporate environment.

From the user perspective, the system should be open, extensible and flexible architecture which supports multiple analysis tools. It needs to support multiple third-party vendors and formats of data (XML, CDISC, SAS, ASCII, EXCEL...). It should be easy to implement and maintain. It must allow users to leverage existing SAS and other proprietary code. Users should be able to work stand alone or within the system. Implementation should be flexible so that specific business rules can be implemented. There should be an ability to create an archive once a study is completed. The actual hardware platforms that the systems run on should be of less concern than the environment that is created.

Productivity depends on alignment of work processes, technology, and organizational structure.

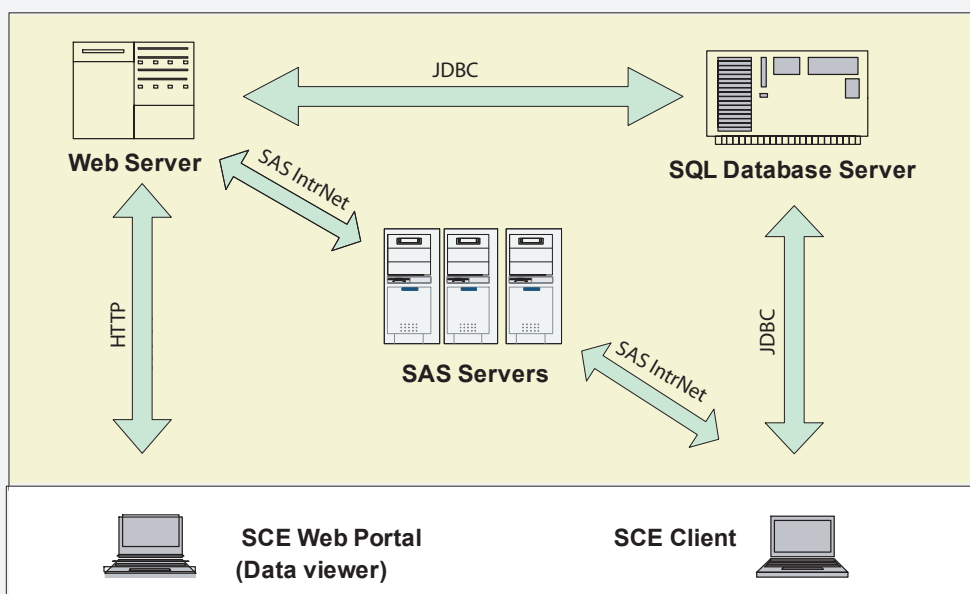
Although software vendors are used to selling “solutions”, technology must work well with people and their roles and work processes. To obtain this balance we believe that the organization should have a strategic vision for technology that is driven by business needs. The strategy should be written and developed by the business process owners. In the clinical-regulatory realm we call this strategy a data vision or data blueprint which focuses on gathering, managing, aggregating and reporting information that is critical to the clinical trials process and regulatory reporting.

- Data Viewer - Access to the data repository**
- Web interface
 - Ability to view raw data
 - Access to TFLs
 - Submit data queries

SCE Advantages:

- Productivity
- Compliance with Part 11
- Facilitates Training
- Manage:
 - Analysis files
 - Programs
 - Reports
- Process documentation
- Audit trail
- File security
- Version control
- Study archival

Figure 2. SCE Technical Architecture



Note: JDBC = Java Database Connectivity, an interface for SQL databases

Systems should be aligned to achieve data transparency (or data traceability) through a simple, integrated, well-documented workflow. The data vision should be supplemented by a technology plan that focuses on integration of the components. Actively review both plans to effectively manage your clinical data pathway.

Summary

The statistical integrity of data analyses and the reporting of results should be a top priority. Establishing a statistical computing environment has several advantages. Tools are targeted directly to the deliverables (clinical study reports, analysis files, eCTD, etc) necessary for regulatory submissions. It organizes the various activities, breaking down the whole process into smaller, simpler tasks. The whole process becomes transparent. It is easier to train new workers and to track progress on large projects. A shared platform usable by both programmers and statisticians facilitates communication and productivity of all concerned. Documentation is created by a structured approach. Finally the environment can create a process that makes Part 11 compliance a by-product of work not an object of work. So an SCE gives a foundation for documenting rigor in the analysis and reporting of clinical trial results while increasing productivity and quality.

Getting Started

PharmaStat recommends the following steps to establishing a statistical computing environment:

1. Create organizational readiness - executive management needs to demonstrate understanding and commitment to the data process as a key to product development success realizing that costs will be incurred and change management will be required.
2. Manage the entire data pathway as one process by developing a comprehensive data vision. The data vision should describe the preferred way of doing business which identifies key roles and responsibilities for those involved in the data process; includes tactical commitments to data standards, data integrity, and other operational standards. One executive should be accountable for the overall process.
3. Develop user requirements for your statistical computing environment needs including validation and quality control of reports.
4. Create a technology plan which indicates how the components of the data vision are managed.
5. Implement in phases, demonstrating return on investment as implementation progresses
6. Align training and standard operating procedures with new systems

About PharmaStat

PharmaStat was founded in 2001 by Alan Hopkins, Ph.D. PharmaStat believes that successful drug development relies on successfully integrating science, technology and clinical operations. The technology and process recommendations made here will be successful only in the presence of a good clinical development program based on good science and appropriate regulatory considerations.

PharmaStat is a service mark of PharmaStat LLC. Other brand and product names are trademarks of their respective companies.

Viewpoint

“The statistical integrity of data analyses and the reporting of results should be a top priority.”

“Manage the entire data pathway as one process.”

PharmaStat LLC
PO Box 598
San Francisco, CA 94104-0598

Contact: Alan Hopkins, Ph.D.
Telephone: 650-598-9860
Fax: 650-598-9861
E-mail: ahopkins@pharmastat.com
Web: www.pharmastat.com