

## Computing Infrastructure

### Information Technology Environment

The Statistical Center computing resources are based on the Microsoft Windows 2000/2003 network. The key services include the patient databases, E-mail, WEB, application and file sharing, electronic document and image management, Cardiff Teleform<sup>®</sup> data forms design and submission, batch application processing, disaster recovery, virus protection services, Citrix-based terminal services, remote access, desktop configuration and network monitoring/management.

### Desktop Systems and Network

Each staff at FHCRC and CRAB has an Intel Pentium III or higher system. Each desktop PC runs MS Windows 2000 Professional or MS Windows XP Professional. Over 35 desktop applications are supported. All workstations access the Statistical Center servers and the Internet. The Statistical Center database, WEB, file and related services are housed at CRAB. Statistical Center staff at FHCRC access CRAB resources (over a dedicated network connection to improve performance and security) using Citrix terminal services and Internet Explorer. Sensitive information is transmitted using encryption. In addition to the dedicated network link between CRAB and FHCRC, each organization has independent connection to the Internet.

### MS Windows Servers

The Southwest Oncology Group Statistical Center uses a modified, distributed architecture with the majority of its MS 2000/2003 Servers, i.e. different servers perform different functions. There are over 50 network servers including a small number of development and test servers. Production servers are Pentium III or higher servers with multiple processors, a minimum of 1 Gigabyte (GB) of memory, have fault tolerant disk subsystems and carry on-site maintenance. Several tape library systems are used for backup.

Key network server services offered at the Statistical Center include:

- **Oracle<sup>®</sup>** Database
- **Microsoft<sup>®</sup>** SQL Server
- **Microsoft<sup>®</sup>** Exchange Post Office
- **Microsoft<sup>®</sup>** IIS WEB Servers (Internet and Intranet)
- **Lyris<sup>™</sup>** Listserv
- **Ecora<sup>©</sup>** Auditor
- **TNT<sup>©</sup>** ELM Event Monitoring
- **Cardiff<sup>™</sup>** Teleform<sup>®</sup>
- **Citrix<sup>®</sup>** Terminal Services
- **Camellia<sup>©</sup>** C/S Batch Services
- **Veritas<sup>™</sup>** Backup Exec Enterprise
- **Trend Micro<sup>™</sup>** Virus Protection
- **Symantec<sup>™</sup>** Ghost Enterprise and WinInstall
- **Microsoft<sup>®</sup>** System Management Services
- **SAS<sup>®</sup>** and **Splus<sup>®</sup>** Statistical Packages

## Data Security and Disaster Recovery

Data Security and Disaster Recovery is based on what are often referred to as “best practices” in electronic computing and networking. CRAB Network Administrators periodically review and compare current network and security best practices with existing CRAB policies and procedures. Outside professional reviews and audits also provide critical information. Updates to policies, procedures and training are incorporated as appropriate.

A summary of key Statistical Center network policies and procedures requirements are:

### Contingency Plans

All servers are backed up to tape (full and incremental). Backup media are transported to and from a bonded, secure off-site storage facility (vault) on a daily and weekly basis. All tapes stored on-site are located in fireproof cabinets in restricted areas. The most recent, and critical backups are stored in fireproof safes with combination locks. Original software, other media and directions are stored in centralized fireproof cabinets.

Servers are configured using fault tolerant disk subsystems. Special disk imaging and file recovery applications are used to speed up restoration on key file servers including database systems. Not only has this process been tested, it is used as a part of normal server construction and upgrades.

Emergency mode plans cover varying levels of disaster recovery developed to address the severity and extent of disaster. This may include a combination of manual and electronic replacement systems until such time critical network services can be re-instated to a fully working level.

### Information Access Control

Supervisors submit Employee Action Forms (EAF) to key account management staff for all employee hires, terminations, or job function changes for staff. User accounts are **user-based**. Individually identifiable patient data, collected only at the time of registration, are stored in a secure table where read and write access is highly restricted and requires prior approval of the Group Statistician, based on demonstration of the user's "need to know" to perform their job functions. All accounts have controlled access to resources based on individual user account definitions.

The highest-level account for network administration is renamed/changed periodically to increase security. For network and database accounts, all staff are required to use “strong” passwords (those with a combination of lower and upper case alphabetic characters, numbers and special characters) to reduce the risk of password cracking. Passwords are aged and staff is required to change them on a regular basis. Passwords are not transmitted via E-mail or over the Internet.

The CRAB Firewalls provide restricted access to and from resources and are monitored 7 X 24. Verisign® Secure Socket Layer (SSL) encryption is used on Web and E-mail servers. Secure File Transfer Protocol (SFTP) is used for file transfer with remote clients.

### Information Technology Security and Monitoring

The CRAB Security and Disaster Recovery Handbook along with the Information Technology Standard Operating Procedures define overall electronic security policies and procedures for the Statistical Center Resources.

Senior level Information Technology staff at CRAB are responsible for monitoring and addressing network security and host/server resources, including the patient database. All Information Technology support staff have appropriate levels of supervision.

Network administrators review network and server event, security and application logs and other reports on a daily basis to monitor login, file access, security incidents and the status of hardware and services. Notification software is configured to provide immediate notification (paging and E-mail) to network administrators for unexpected network and server events.

Servers and workstations are proactively updated with security patches as well as OS and application updates. Network Administrators subscribe to notification lists to stay on top of emerging problems and corresponding updates or fixes.

All desktop computers, servers and the E-mail post office have active, real-time virus protection. Virus protection software is automatically pushed to the CRAB network and subsequently updated on each system.

Computer equipment (including equipment checked out to staff) is logged and tracked in an online inventory database including location and user (for individual desktops).

Identified real or suspected security incidents are logged, addressed and reported to and/or by senior network administrators and the Director of Information Technology. Problems are in turn reported and/or escalated to CRAB Executive Offices and the CRAB Professional Conduct Board as appropriate.

Senior management including senior network administrators perform risk assessment on new systems and events to define the cost and benefits of different solutions and the solution impact on: Confidentiality, Integrity and Accessibility.

### Media Controls

All software media and licensing are filed in fireproof file cabinets and restricted areas. Only authorized personnel may access original software media and licensing. All software upgrades to workstations, servers or other systems are done by Information Technology staff or by explicit permission of the Information Technology Director (in very rare situations, some users may install specified software licensing).

All on-site backup media are stored in either a fireproof safe or fireproof cabinet located in restricted access areas where only authorized staff may enter. All off-site backup media are stored in a secure, bonded, protected vault at professional facilities.

All old server and PC disk drives, CDs, portable media and other storage material are destroyed by a bonded, professional media destruction company.

All server rooms, server configuration areas and media storage are in restricted access locations. Signs are clearly posted on restricted areas noting that they are restricted. Only a

limited number of authorized staff may enter restricted areas. All server rooms and server configuration areas have security cameras, which record 7 X 24 hours per week. All servers have screen/keyboard locks. All vendors or other third party visitors accessing restricted areas are logged into an electronic file and escorted by authorized staff.

### Information Technology Resource Access Policies

All staff are required to read and sign software, network, computer and data protection policies. These policies clearly outline and define proper use of desktop computers, E-mail, access to other network services, software usage, protection of patient information and other related practices. Operating System policies may further restrict staff from inappropriate computer or software usage. Network software monitoring applications track software use by user and workstation. All users must have valid accounts and passwords. Logon sessions have enforced password protected screen savers that lock the systems after 30 minutes of inactivity. Staff workstations are located in specific work areas that are locked after normal business hours. Remote and local users must maintain integrity and confidentiality when accessing CRAB resources.

### Data Authentication and Encryption

The use of encryption (VPN, SSL and SFTP) reduces the risk of alteration or easily viewing of Internet traffic (packets) containing sensitive information. Operating system and database controls restrict inappropriate access privileges to data, files and other objects that require protection from modification.

## **Software**

The operation of the Southwest Oncology Group Statistical Center depends on six major classes of software: database management, statistical analysis, desktop applications, network services, report processing, and data management.

### Database Management

The database management software used is Oracle, one of the major commercially marketed systems. Oracle is based on the relational model of database management and is built around the industry-standard SQL language. The Statistical Center's data management operation is built around Oracle's capability for multiple users to manage simultaneous database modifications. In addition to the core relational database management module, Oracle has components for ad hoc queries, report writing, generation of screen-based data maintenance applications, interfacing to high-level languages (C++ or Visual Basic, for example), and database administration and tuning.

### Statistical Analysis

The main statistical package used is SAS™. Several in-house programs have been written using SAS™ and Splus™ to perform tasks such as Cox regression diagnostics, Kaplan-Meier survival curves, sample size computations, exact methods, recursive partitioning, and longitudinal data analysis. The Center has recently acquired licenses for GenePlus and Insightful ArrayAnalyzer software for the analysis of microarray data.

### Desktop Applications

Several Microsoft desktop applications are used by staff including Office Professional (Word, Excel, PowerPoint and Access), Visio and Project. Another key application is Adobe Acrobat.

### Network Services

Electronic mail is used extensively for communication within the Statistical Center as well as with the Operations Office and with other Group members. CRAB and FHCRC staff access respective Microsoft Exchange Post Offices with MS Outlook and Outlook Web Access (OWA). The Statistical Center uses MS Internet Information Server (IIS) for Web services. Cardiff Teleform® is used for forms design, data submission and data entry.

The Southwest Oncology Group Home page can be found at <http://swog.org>. This site is maintained at the Operations Office in San Antonio. The swog.org website has links to the Statistical Center's services.

### Report Processing

The Statistical Center Report of Studies is created using an application developed at the Statistical Center (Statisticians' Report Worksheet, or SRW) incorporating a web-based interface, creation of a SAS data set from Oracle, and the word processing tools of Microsoft Word.

SRW is based on a "thin" Client/Server (C/S) model using Web publishing technology. Web pages are driven from two primary database sources: Oracle and SAS data sets. MS Internet Explorer provides a program interface for input of textual and study parameters needed to define and set up charts, tables, graphs, and descriptive information. SAS extracts the data from the patient database via Open Database Connectivity to create a SAS data set, i.e., a "snapshot" of the patient data. Study chapter generation is done on a Web server based on input from the SAS data sets, study information defined in the Oracle database and end user input (such as text, label definitions, and table format information).

Users are able to view and output the study/chapter results in three ways:

1. As Web pages for preliminary browsing/viewing of the document.
2. As Web pages for final publication.
3. As formal output to a printer, a postscript file, and other file format, e.g., MS Word or Portable Data Format (PDF), for professional printing of the Report of Studies. PDF copies of the Report of Studies are available on the SWOG Home page (<http://swog.org>).

Hyper Text Markup Language (HTML) templates represent the various tables/charts. Microsoft's IIS Web Server and some of its key components are used to provide Web and final publication. A program process fills in much of the templates based on the SAS data sets and other stored database information, in order to generate more complete HTML

documents. Additional programming filters further refine the Web pages, and the formal output of the Report of Studies (ROS) is an executed "object linking and embedded-enabled document production manager", as the final ROS requires more extensive formatting (headers, footers, page numbers) compared to the Web pages.

The study chapters are made available as static publications based on the "snapshot" data sets. Data sets, other interim priority documents and the final chapter output are archived for future retrieval and reference. Since this model is a mix of "thin" C/S and Web publication with some portions being batched off to back-end Windows servers, it works well for both Local Area Network and Remote Access clients.

Security is built into the systems: correct user account information and password protection are required for accessing these services and firewalls are used to audit and restrict access.

Other in-house reports are created using SQL queries and Microsoft Active Server Pages (ASP) technologies to provide dynamic links to the Oracle database. Production reports and other documents are created using Microsoft Word, PCTex and Scientific Workplace.

### Data Management

The main data management systems have been written in-house and consist of the patient/participant registration and randomization system (WebReg), the patient evaluation system (EVE), and Chart Manager, for the creation, manipulation and viewing of electronic patient charts. These programs all run on Windows, and are being re-written in .NET.

### *WebReg (Patient Registration)*

The patient registration and randomization program is very complex, since it is a generalized program that facilitates study set-up and patient registration. It is designed to handle the wide variety of requirements for SWOG studies. It is a thin client application designed to be used in-house or via the Internet by a CRA. It is written in HTML with one C++ routine to interface with the Oracle database. Where this system is unavailable, registrations are also performed by telephoning a data coordinator in the Statistical Center who is online to the WebReg program. Randomization is typically accomplished using a dynamic balancing algorithm, which uses current accrual counts by stratification variables directly from the database. The Southwest Oncology Group Web registration program is also able to perform direct Internet patient registrations to intergroup studies coordinated by ECOG.

### *EVE (Patient Evaluation System)*

The patient evaluation system allows data coordinators to update patient data such as toxicity, response, and vital status. Its design also allows the capture of data unique to a particular protocol. This system is written in Visual Basic and runs on Windows. A major function of EVE is to provide cross-field edit checks for evaluation data fields.

### *Chart Manager (Image Management)*

Several years ago, the Statistical Center made the decision to move towards electronic data image management for our therapeutic studies. Because we did not want to maintain

both paper and electronic systems into the future, we undertook the scanning of 80,000 paper charts (2.2 million images), now completed. We now create, maintain and view electronic charts using an in-house application (Chart Manager), which is fully integrated with MS Windows and Oracle. All Teleform<sup>®</sup> patient file images and future data entry documents are included in the document management system (as TIFFs).

All SWOG data coordinators have client query tools for viewing patient charts, query and annotation utilities, and redaction (for HIPAA compliance). WebReg and EVE integrate directly with Chart Manager. Future enhancements will include the ability to view electronic data in other than TIFF format, as is being developed for the Selenium and Vitamin E Cancer Prevention Trial (SELECT).

## Database

The Southwest Oncology Group Statistical Center manages its database using Oracle<sup>®</sup>, a relational database management system. There is a production database that stores data that are reflective of real events, and a test database that is used for ongoing development and testing of applications. Each database is organized into two schemas, one for staging tables and one for active tables. Active tables hold data that are included in evaluations and analysis. Staging tables are used to store submission attempts from electronic data entry (EDC) applications, attempts that may violate pre-determined business rules. Once the business rules are cleared, data from the staging tables are promoted to the active tables.

The table structure in each database is organized into five main components:

1. **Common Patient:** This component contains patient-related data items, which are common to all Group studies including patient characteristics; identification of investigators and institutions to which a patient is associated; registration date; stratification; treatment; common evaluation items; and adverse events. Many of these items are accessed frequently in the day-to-day operations of the Statistical Center.
2. **Study Characteristics:** These describe the Group studies and are available to the data operations software, which must modify its behavior depending on the study being processed.
3. **Membership:** The membership component describes the investigators, clinical research associates, institutions, pharmacies, labs, radiation therapy facilities, bone marrow transplant facilities, and the relationships between people and sites. We further describe members by the web site and registration permissions they have for each site affiliation.
4. **Detailed Patient:** Detailed committee or study specific patient data items are regarded as a separate database component. These are not accessed as frequently as the common patient data items.
5. **Quality Control:** A major aspect of quality control data is our generalized tracking system, which stores data Expectations, and tracks submission of the required study information. We also maintain data from various review processes (pathology, surgery and radiation therapy). In addition, the data operations software forces quality control standards and collects quality control data.

Following are descriptions of some of the major tables in the database. The indications of size are given as rounded numbers from June 2004. These descriptions are in relational database management terminology. The relational model can be thought of as organizing data into tables that can be linked to other tables using key variables. For example, patient number is a key variable used to link prestudy data, adverse event data, and patient registration information (all residing in separate tables) for a single patient.

#### PAT table

The PAT table contains data on the patients enrolled into Group studies. At present, PAT contains data on over 160,000 patients. The columns include the following identifiers: patient number, birth date, sex, race/ethnicity, date of last contact or death, vital status and current follow-up investigator and institution. Confidential patient data, such as name and social security number are stored in this table under an extra layer of security, to which only approved personnel have access.

#### REG table

The table REG has one record for each registration. A patient may be registered to more than one study and a study may have more than one registration step. At present, there are over 225,000 registration records in REG. The columns include registration date, the institution and investigator credited with the registration, assigned treatment arm, and indicators for various review requirements (surgical review, for example). In addition, REG is the place where common patient evaluation data are stored for most studies. These include eligibility status indicators, treatment dates, response and relapse data, treatment deviation indicators, adverse event evaluability data, and treatment status indicators.

#### TOX table

The table TOX contains one row for each adverse event reported. At present, TOX has over 500,000 rows. Columns include cycle identifiers, adverse event, severity, and attribution.

#### STUDY table

The table STUDY contains one row for each possible type of registration. At present, STUDY contains data on over 950 studies. Approximately 115 registration types from 88 studies are open to accrual. The columns of STUDY include study and registration identification, study name, activation status, phase, study characteristics, review requirement indicators, and dates opened and closed (if applicable). The table STUDY is part of a large complex of tables, which describe various parts of studies, such as treatment and stratification information.

#### ROSTER table

The table ROSTER consists of one row for each investigator, clinical research associate, affiliate or individual for whom the Statistical Center requires contact information. There are over 23,000 rows. The columns include name, address, phone number, fax number, and email. ROSTER is the central table in a large collection of tables, which provide lists of in-

investigator subsets. For example, table INVEST provides the list of investigators and their current Group affiliations.

EXPECT table

Each row of table EXPECT contains data on an expectation for forms or material submission from an institution for a patient with respect to a particular study. EXPECT is a very large and active table, currently with over 5.8 million rows. The columns include the posting date, the due date and the resolution date.

AUDIT TRAIL table

In October 2000, a comprehensive audit trail was implemented for the entire production database. Every insert, update, and deletion for data from the production database is recorded in the audit trail. The audit trail records the user name, terminal name, date and time of the change, table name, unique identifier for the row, column name, and the old and new value.

Other tables

The detailed patient data component of the database consists of a collection of over 700 tables. Many are study specific, or are applicable to a class of studies (breast cancer studies, for example). These tables generally have many more columns than the tables in the other database components; several have over 200. Many contain prestudy (baseline) data, but others are used for special detailed event data or detailed pathology data.