An Application of the
Internet-based Automated Data Management System (IADMS)
for a Multi-Site Public Health Project

Michele G. Mandel, National Centers for Disease Control and Prevention, Atlanta, GA
Robert E. Schwartz, Jr., National Centers for Disease Control and Prevention, Atlanta, GA
Steven A. Kinchen, National Centers for Disease Control and Prevention, Atlanta, GA

Abstract

In 1992, the National Centers for Disease Control and Prevention (CDC) began a multi-center case-control study of breast cancer risk factors, focusing especially on risks associated with use of oral contraceptives and other steroid hormones. Data collection for the study, which is already in progress, will last four years. The study is a collaboration between the National Institutes of Health, National Institute of Child Health and Human Development (NICHD), CDC, five field sites, and a data management contractor. Initially, data processing relied on diskettes and mainframe tapes to transfer data among field sites, the data management contractor, and CDC. Data were loaded onto the CDC mainframe and processed in a batch environment. The growth of the Internet and the availability of the SAS® System on a UNIX workstation led to a redesign of the data processing protocol and development of the Internet-based Automated Data Management System (IADMS). File Transfer Protocol (FTP), cron scheduling, and shell scripts detect and transfer incoming data to the analysis platform, run reports, and E-mail results to data managers. Using the IADMS eliminated several manual processing steps and automated project reporting, thereby improving data management and staff efficiency. Additionally, the new protocol and platform enable researchers to use advanced interactive analysis tools, such as SAS/INSIGHT®.

Introduction

Public health studies often involve multiple partner organizations located at different geographic locations. The (NICHD) Women's Contraceptive and Reproductive Experiences (CARE) Study is a multi-center case-control study of breast cancer risk factors, focusing especially on risks associated with use of oral contraceptives and other steroid hormones. Data collection for the study, which is already in progress, will last four years. The study is administered by the CDC, National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP) and Division of Reproductive Health (DRH). Five metropolitan area field sites and a data management contractor (DMC) collect, edit, and transmit data monthly to CDC. CDC merges the files, analyzes them for correctness and consistency, and generates project management reports. When the CARE Study began, the DMC used a courier service to transfer data tapes to CDC. This method was expensive, labor intensive, and prone to delays. Further delays were introduced by the number of manual steps necessary to transfer the data from tapes to SAS data sets on the CDC mainframe. Analysts and programmers working on the CARE Study recognized the need for a more efficient and reliable system for transfer and initial processing of study data.

Original Data Management Protocol

Before implementing the data management system described here, the CARE Study used a manual data management process involving staff from several departments. This process, which occurred monthly, began when the cut-off date for a batch of data records was reached. The DMC performed editing and quality control checks on the scientific data supplied by the field sites and then concatenated it to a cumulative file. When this editing was completed, the data were converted to a SAS transport format data library on 9-track tape and sent to the CDC project officer. Project management data were sent to CDC monthly, initially on floppy disks and later by electronic mail attachments.

The project officer logged in the shipment and gave the tape to the CARE Study data manager. The data manager sent the tape via interoffice mail to CDC's mainframe data center, which is located at a separate facility in Atlanta. One to two days later, the transport library tape was available to be read into native SAS data libraries using the SAS System in batch mode. After the data were copied from the tape, the tape was mailed back to the data manager for return to the DMC. At this point, the data were ready to be used for project management reports and scientific analysis.

The initial data management protocol for the study was designed in 1991, starting with an assessment of available hardware and software. Access to the SAS System on a UNIX platform was unavailable at that time. The SAS System for personal computers running on Intel-based computers was considered. An adequate PC configuration would have cost about $7,000 per seat. A shared laser printer added an additional $1500 to the cost. Five staff members needed to be equipped to work with the data, which would have brought the equipment cost to an unacceptably high cost of $36,500. Another drawback of the PC platform was that it did not have the robust batch job processing and file security facilities of other platforms. CDC's mainframe was chosen instead because of staff familiarity with its use and its favorable cost. Mainframe computing is centrally funded at CDC; individual projects are not charged directly for use. Electronic data transmission issues were not considered at that time.

Drawbacks of the Original Data Management Protocol

The CARE Study has been collecting data for two years. During this time a number of shortcomings became increasingly apparent in the original data management protocol. The most substantial of these is the shipment processing time. The following table shows approximate times for each of the major steps in the original protocol.

Task Duration for Original Protocol in Days	
Data tape preparation by contractor	1
Transfer tape via a courier to CDC	2
Log in shipment and send to CDC data center via interoffice mail	2
Run mainframe SAS jobs	1
Return tape from the data center via interoffice mail	2
Return tape to the DMC	2
Total	**10**

After data editing at the DMC it took six days to prepare the data for use. Another four days were needed to return the tape to the DMC. Once the data tape was at the data center, the project programmer spent most of one day running batch jobs to prepare the data for analysis.

The mainframe computing environment itself has become an issue during the past six years. In 1991 a working knowledge of the CDC mainframe environment was required to use the SAS System. Today many scientists and statisticians use the SAS System for Windows as their primary platform. Newly hired staff are far less likely to have extensive mainframe computing experience today than they were in 1991. The need to pre-allocate adequate library space is unique to the mainframe in the CDC environment. Even regular mainframe users are not always skilled in estimating space requirements. This results in repeated executions of jobs, and increased user frustration. The SAS System now contains powerful interactive facilities for exploratory data analysis (SAS/ASSIST® and SAS/INSIGHT). These facilities are available on the mainframe, but they do not function optimally in a terminal emulation environment. Finally, job turnaround time on the mainframe is variable, with queue delays of 40 minutes or more possible during peak periods. All of these factors detract from using the mainframe as a primary project platform.

Design of the IADMS

To improve processing efficiency, reduce the number of steps in the data transfer process, and address resource issues, DRH collaborated with the NCCDPHP Information Resource Management (IRM) Activity to design an efficient, secure, and automated data processing system for the CARE Study. An analysis of the project data flow was performed, with consideration given to changes that have taken place in the CDC computing environment since 1991. The Internet, which was not widely available in 1991, can now be placed on every desktop. NCCDPHP and other centers within CDC have purchased scientific workstations and servers running the UNIX operating system to support data analysis using the SAS System. A secure FTP server was established in 1995 to support data communication with field collaborators. These enhancements in the computing environment permitted the design of a new model for the CARE Study data flow. The result is the Internet-based Automated Data Management System (IADMS).

As shown in figure 1, the IADMS has introduced several new components to the data flow of the CARE Study. The field sites and the DMC transmit data electronically to CDC via the Internet. Their data transmissions are placed on the CDC secure FTP server in private directories. Programs on the statistical server detect the presence of new data on the FTP server, move the data behind the firewall, and generate summary reports for project managers.

Internet FTP Reduces Transfer Times

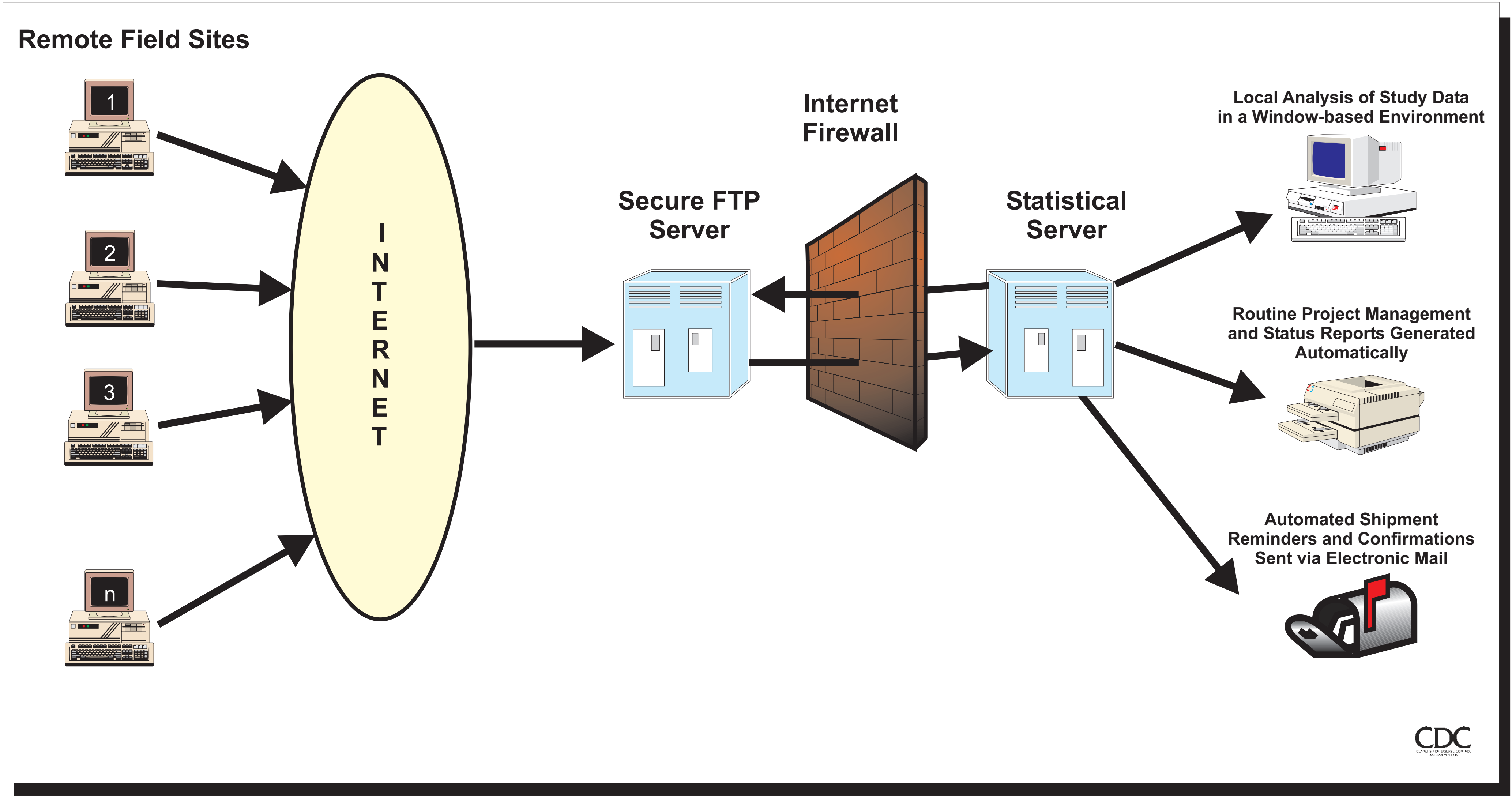
The review of the data management protocol for the study identified one major inefficiency. Using couriers to transfer tape media accounted for 66% of the time needed to process a data shipment. Current technology permits a faster data transfer solution. All of the field sites and the DMC were queried and found to have Internet services available. This enabled the study to use the Internet FTP service for routine data transfers with no added expense.

The NCCDPHP statistical server has FTP capability for transferring data in-house between different host and client platforms. This enables researchers to move data easily between UNIX-based platforms, PC's, and the CDC mainframe. For security reasons, the CDC Internet firewall blocks inbound FTP service requests from outside the organization. The CDC secure FTP server was established outside the firewall to enable collaborators in other organizations to transmit data electronically. The firewall permits computers within CDC to access the secure FTP server.

Each field site has a separate account on the secure FTP server. These accounts allow field sites to transfer data sets to a private directory on the server and to read files from a shared resource directory. CDC project staff members have a management account that has full rights to all of the field site directories and the shared resource directory.

File compression improves the efficiency of data transfers. SAS data sets containing scientific study data are converted to transport format and compressed with the GNU Software Foundation utility GZIP. The PKZIP file compression utility is used to compress multiple project management data tables into a single archive file for transmission. These compressed files are transferred by the field sites and the DMC to the secure FTP server monthly. This process replaces the sending of mainframe data tapes, PC floppy disks, and electronic mail attachments.

Figure 1


Remote Field Sites
1
2
3
n
INTERNET
Secure FTP Server
Internet Firewall
Statistical Server
Local Analysis of Study Data in a Window-based Environment
Routine Project Management and Status Reports Generated Automatically
Automated Shipment Reminders and Confirmations Sent via Electronic Mail
CDC

UNIX Process Scheduling Automates Routine Tasks

Incorporating FTP into the data management protocol allows CDC to access study data the same day it is sent. However, the project programmer is faced with a new series of routine tasks necessary to detect, retrieve, and process data shipments. The UNIX task scheduler CRON simplifies these tasks for project programmers so they can focus on new software development.

CRON is one of the automated system management facilities available under the UNIX operating system. One of the primary uses of CRON is to run file system backups at scheduled times. While CRON is most often used as a system management tool, it is also available to other users for scheduling routine tasks. Scheduling tasks for execution by CRON is a straightforward process. A text file is created that contains one line for each scheduled task. This file is registered with CRON by running the Crontab command, as shown below.

`crontab` ***`filename`***

A task is the execution of a compiled program or a shell script. A shell script is a series of commands stored in an ASCII text file. Each line in a crontab file contains six fields, separated by spaces or tabs. The first five fields specify the execution time(s) for the task. These fields contain either an asterisk (meaning all legal values) or a comma-delimited list of elements that specify the following:

```
Field 1:  minute (0-59)
Field 2:  hour (0-23)
Field 3:  day of month (1-31)
Field 4:  month of year (1-12)
Field 5:  day of week (0-6 with 0=Sunday)
```

Elements can be either single values, or two values separated by a minus sign (meaning an inclusive range.) The remainder of the line contains the command to be executed. An excerpt from the Crontab developed for the CARE Study is as follows:

```
0 12 * * * csh /sasmt/home/care/sampid.shl
0 12 * * * csh /sasmt/home/care/catidata.shl
0 12 * * * csh /sasmt/home/care/cumcare.shl
0 12 * * * csh /sasmt/home/care/caretrac.shl
0 6 30 1 * csh /sasmt/home/care/caremsg.shl
0 6 28 2 * csh /sasmt/home/care/caremsg.shl
                        •
                        •
                        •
0 6 27 11 * csh /sasmt/home/care/caremsg.shl
0 6 30 12 * csh /sasmt/home/care/caremsg.shl
```

Processes Performed on Statistical Server

This crontab is run on the statistical server. The first four tasks in the schedule automate the detection of new data transmissions. Each task processes a different type of study data. The data are moved across the firewall to the statistical server, and content-specific processing is performed. These tasks are run on a daily basis to allow the field sites and the DMC some scheduling flexibility in preparing and sending data. The remaining tasks in the schedule send E-mail messages to the field sites monthly, to remind them to FTP their project management data to CDC.

Each data processing task is implemented as a C-shell script. The C-shell is one of the command interpreters available in the UNIX environment. The script ***`cumcare.shl`*** processes the scientific study data transmitted from the DMC to CDC. This script is as follows:

```
source /master_cshrc
cd /sasmt/home/care
ftp -n < get_cumcare
if ( -e "cumcare.xpz" ) then
    rm cumcare.xpt
    gzip -dN -S .xpz cumcare
    sas sasexpq.sas
    sas cummerge.sas
    sas carefqfm.sas
    sas carewarn.sas
    mail uuuu@cdc.gov < freq.msg
    mail vvvv@cdc.gov < cumcare.msg
    mail wwww@cdc.gov < cumcare.msg
    mail xxxx@cdc.gov < cumcare.msg
    mail yyyy@cdc.gov < sasexpq.log
    mail zzzz@cdc.gov < backup.msg
endif
```

Script execution begins by performing site specific initialization and setting the current directory to the project work directory. The command line FTP program is then run, processing commands stored in the file ***`get_cumcare`***. These commands connect to the secure FTP server and attempt to retrieve the uploaded data file ***`cumcare.xpz`***. If the uploaded file does not exist, the script exits immediately. When the uploaded file is found, it is copied to the statistical server and renamed on the secure FTP server to prevent future checks from finding it again. The previous shipment transport file is removed, and the new uploaded file is decompressed using GZIP. The resulting file, ***`cumcare.xpt`***, is a SAS transport format data library containing five SAS data sets. These data sets are converted to native format, merged into an analysis data set, and a series of univariate tables and diagnostic reports are produced. Finally, E-mail messages are sent to various project team members informing them of the shipment status.

Results

The IADMS has provided many benefits to the CARE Study. Data processing efficiency has been significantly improved by cutting an average of four days from the turnaround time required to prepare data shipments for use, and by eliminating the extra four days and shipment step needed to return used tapes to the DMC. Project communications are more timely because the system notifies field sites of data receipt and provides comprehensive edit reports to appropriate staff within one day of data transmission.

The following table shows the time saved using the IADMS to relay study data.

Timeliness Comparison

Task	Previous System	IADMS
Data tape preparation by contractor (Includes FTP under IADMS)	1	1
Transfer via courier to CDC	2	NA
Log tape in and interoffice mail to CDC data center	2	NA
Mainframe SAS jobs (SAS processing is performed in the UNIX environment under IADMS)	1	1
Return of tape from data center via interoffice mail:	2	NA
Return tape to DMC	2	NA
Total	**10**	**2**

Data quality has been improved because the standardized programs used in the IADMS ensure consistent, precise processing procedures and that improvements made to any program component will be used on all data. Data security has been enhanced by reducing the number of manual steps and physical transfers of the data and by ensuring the timely processing of submitted files. Finally, the IADMS has eliminated the need for a programmer to manually execute processing jobs. This enables the programmer time to be allocated to tasks such as designing new reports and analyses for automation by the IADMS. During the remaining two years of data collection, the CARE Study will achieve substantial time savings using the IADMS.

Future Plans: CARE data managers will extend the number and type of SAS programs that the IADMS runs. In addition to basic file processing, data managers have added programs to automatically generate certain subsets of the CARE Study data. The IADMS puts these files on the FTP server and informs the field sites of their availability by E-mail. DRH and other divisions in NCCDPHP plan to use the IADMS for additional studies and data collection projects.

Conclusion

The IADMS has improved data processing efficiency, data quality, security and collaborator communication for a multi-site, multi-partner research project while simultaneously reducing staff time requirements. It could provide the similar benefits to other data collection efforts.

Its main strengths include:

- *Reliance on industry standard transport mechanisms:* The Internet is a globally supported standards-based system. This allows multiple organizations to link data collection and processing systems easily and reliably. By using the Internet for transport, IADMS will be able to take advantage of improvements in transport technology as they become available.

- *Reliance on widely available and well-known software tools:* The IADMS relies on the SAS System and standard UNIX operating system functions and utilities. Any UNIX environment running the SAS System would be able to implement a similar system and achieve similar results.

- *Scalability:* The IADMS could be used in studies that differ in data set size, data set complexity, number of sites, type of hardware, and timeliness requirements. Because UNIX is scalable, any number of sites and connections could be supported with the same methodology; no special programming or techniques would be needed to accommodate a data collection and processing project of several hundred sites. In fact, data collection for numerous sites would require a comparable degree of automation.

Using the IADMS successfully in a study also requires careful consideration of several issues:

- *Data security:* The Internet does not yet support fully secure transmissions although that is improving. Therefore, it is necessary to design checks into the data processing routines to ensure that file transmissions occur correctly. In the CARE Study, record counts and edit reports are carefully reviewed to ensure correct transmission.

- *Confidentiality:* The CARE study does not collect any information that could be used to identify a specific study participant. Any study that collected personal identifiers or sensitive information of any kind would have to apply methods to ensure that data confidentiality was maintained during transmission and processing. Encryption and file splitting should be considered and conscientiously applied.

- *Internet Access:* All study collaborators must have Internet access with sufficient reliability and communications bandwidth to support the intended levels of data transmission. High speed modems or data connections will be required as well as local support to assist field sites if problems arise.

- *Host Throughput:* Sufficient host resources are needed to process projected FTP traffic, perform scheduled program tasks, and continue to meet the needs of other system users. If performance is a concern, and timeliness restraints permit, scheduling the IADMS SAS jobs to run at off peak hours to spread the load over a longer period of time would be possible but would impact the advantages that come from immediate data set turnaround.

References

SAS Companion for UNIX Environments: Language, Version 6, First Edition. SAS Institute Inc., Cary, NC.

UNIX System Documentation

PROTOCOL for The NICHD Women's Contraceptive and Reproductive Experiences (CARE) Study: A Case-Control Study of Breast Cancer and Hormone Use

Authors

Michele G. Mandel
Division of Reproductive Health
National Center for Chronic Disease Prevention
and Health Promotion, CDC
4770 Buford Hwy, N.E. MS K-21
Atlanta, GA 30341-3724
E-mail: mgm2@cdc.gov

Robert E. Schwartz, Jr.
Division of Reproductive Health
National Center for Chronic Disease Prevention
and Health Promotion, CDC
4770 Buford Hwy, N.E. MS K-21
Atlanta, GA 30341-3724
E-mail: res1@cdc.gov

Steven A. Kinchen
Division of Adolescent and School Health
National Center for Chronic Disease Prevention
and Health Promotion, CDC
4770 Buford Hwy, N.E. MS K-33
Atlanta, GA 30341-3724
E-mail: sak1@cdc.gov

Acknowledgements

The authors express their appreciation to the staff of the Women's CARE Study and to Margaret Watson for their continued encouragement and support of this project.