Abstract

Merck has submitted their second generation Statistical Computer Assisted New Drug Application (CANDA) to the Food and Drug Administration (FDA). Continued development at Merck of the CANDA focused on the implementation of advanced features, analysis and implementation of a new hardware platform and the initial incorporation of the CANDA as an internal New Drug Application (NDA) development tool.

Introduction

A Computer Assisted NDA system gives the FDA a welcome set of tools for the reviewer responsible for evaluating the paper version of the NDA. The computerized system provides the enhanced ability for the FDA to study the reasoning and methodologies implemented by the company during the clinical trial process and to better understand the effects of the chemical or biological entity on a cross section of a targeted population.

The CANDAs submitted in the past, have performed with varying degrees of success at the agency. The more creative the implementations of the CANDA, the broader the exposure the agency gains in assessing the features and functionalities that can assist them in their evaluation of the drug entity.

For the second generation of CANDAs at Merck, we applied our previous CANDA experience, FDA comments stated both publicly and privately to us, and computer technology available at the current time of development.

Hardware

Merck’s first CANDA was developed on the IBM 3090 mainframe. The FDA had requested future CANDAs to be submitted on personal computers (PCs) due to telecommunication problems associated with accessing a sponsors computer.

We began development of the second generation CANDA on an IBM 486/25 MHz machine using DOS v3.3. The PC had 4 Mb of memory and 300 Mb of hard disk storage. Utilizing Quarterdeck’s Expanded Memory Manager v5.0, the four Mb of memory were configured with two Mb for extended memory and two Mb for expanded memory. The 2 Mb for expanded memory were optimal for SAS® software because only the executable Procs load into expanded memory. The device drivers, DOS buffers, the mouse and ROM were also loaded into expanded memory. The execution of large I/O intensive SAS programs with SAS input data files up to 150,000 records with fifteen variables and SAS/AF window overhead, produced less than desirable results under the DOS operating system. Extremely poor response time coupled with the constant shortage of memory led us to investigate the OS/2 operating system.

The OS/2 operating system did solve the memory problems. However, there was still concern over the efficiency of the SAS programs and the resulting end user response time. To address these concerns we investigated the new capabilities of version six SAS software. Table 1 lists benchmark results using two new features within version 6.06; the ability to index the SAS data set and the usage of the ‘Where’ statement. The usage of the Where statement on an indexed file produced tremendous savings in processing time. The PC SAS programmer must implement the most efficient code to process the data or else unproductive waiting time for the end user will occur. The longest running statistical program tested executed in approximately ninety (90) minutes.

It was soon after the move to the OS/2 operating system that the company decided to substantially increase the amount of safety and efficacy data in the submission. This clearly warranted another look at the proposed hardware for this generation of the CANDA. There were only two choices for the new hardware platform, back on the company’s mainframe or a personal workstation. Since the mainframe was not a desirable platform for the FDA, an evaluation of the workstation environment was initiated. After much consideration of the four prominent workstation manufacturers, the Digital Corporation’s workstation system was chosen. The reasons for this choice at this point in time were the critical availability of the equipment, the availability of the Ultrix version of SAS 6.07 software and the technical support provided by Digital on the project.

The second generation CANDA development was completed on a DECStation 5000 model 200 workstation. The workstation contained 32 Mb of memory, 1.655 Gb of hard disk storage, a floppy drive, a 95 Mb tape backup system, a nineteen inch high resolution monitor and a laser printer.

Software

Production versions of SAS 6.07 products for the RISC/Ultrix platform were utilized for this application: SAS Basics, SAS/AF, SAS/FSP, SAS/GRAPH, SAS/STAT, and SAS/INSIGHT.
Other software included the Ultrix operating system, SoftPC to create a DOS shell, and WordPerfect.

**Methods**

In order to integrate the Statistical CANDA with the concurrent development of the paper NDA and to minimize the introduction of errors into the final product, utilization and expansion of computerized tools already required to produce the paper NDA were necessary. The SAS programs used to produce the Case Report Form (CRF) Tabulations located in Item 11 of the NDA and SAS statistical programs used to produce the efficacy results in the two pivotal Clinical Study Reports were cornerstones of the CANDA implementation.

The safety data for the CANDA, the safety tables contained within the Clinical Study Report and the CRF Tabulations were extracted from the same SAS program (figure 1).

By expanding the program after the Data_Null_step, that has historically only produced Merck's CRF Tabulations, the program created a hardcopy of the Clinical Study Report (CSR) safety tables and electronic versions for both the CANDA and the word processing software used to produce the paper NDA. This proved to be a successful way to automate the extraction of the safety data and a method to incorporate CANDA activities into the daily routines of producing the paper NDA.

The efficacy data were handled in a similar fashion. The data flow used to produce the efficacy results satisfied both internal needs to produce the paper NDA and the needs of the CANDA. Consultations with in-house statisticians and with FDA statisticians enabled the efficacy programs written to meet both groups needs with minimal additional effort. The data variables to be contained within the CANDA application were subsequently agreed upon by the FDA.

During our consultations, the FDA statisticians requested all initial, intermediate, and resultant data sets, programs, tables and graphs associated with an analysis to be available within the CANDA. The FDA reviewed the level of documentation in a sample SAS program and requested that we drastically reduce macro code due to its complexity. They also asked that we implement patient compliance flags to state what population the patient belonged, the Intention-to-Treat and/or the Per-Protocol population. The FDA asked to start the CANDA at the extraction of the data from our Inquire Clinical Database Management System (figure 2). In order for us to create a more streamlined and efficient system design, the FDA agreed to our proposal to start the CANDA after the data had been reorganized from Merck's database into a more easily understood SAS dataset and the relative day from study start was added to the data. Before any analysis of the data, a program that reorganized the data for the statistical analysis programs was needed. This program was known as the Set-Up program. The Set-Up program was responsible for grouping relative days into common time points, assigning "patient compliance codes" to the population, estimating missing data where applicable, and performing the appropriate data manipulation for the analysis programs.

Merck's implementation of the FDA's request to associate an efficacy flag to patient observations was substantial and complicated for this investigated new drug. We
created patient compliance codes and decodes to associate every efficacy observation with how the patient was or was not violating the company’s approved protocol. Each unique description was assigned a number which was the first digit of this four digit code. Extensions of the patient compliance code request were achieved by adding digits to the left of the patient compliance code. By using binary flag fields (0 or 1) in the other positions of the patient compliance code, criteria specific to important conditions associated with the drug were indicated. In addition, forming the usual pre-specified patient populations (intention-to-treat, per protocol) became simple functions of the codes. The Set-Up program contained all the SAS code used to assign these patient compliance codes. If patient compliance codes were assigned to more than one efficacy variable then verification between Set-Up programs was needed to ensure consistent assignments of the codes to the patients. Having this logic available to the FDA, allowed them to add, change or delete the codes and definitions as needed.

The in-house utilization of patient compliance codes created a more consistent, unified and manageable approach dealing with subsetting large patient populations. The selection of patient compliance codes created a population that was then analyzed with a single generic program that could handle many different efficacy parameters. The statistical patient logs, also part of the paper NDA, which were very time consuming for the statistician to maintain were now a fallout of the implementation of the patient compliance codes.

Once the data proceeded through the Set-Up program, implementing the statistical analyses was straightforward. The Analysis/Graphic program was responsible for the definition and execution of the statistical analysis, the report formats displayed in the body of the CSR, the expanded table located in the appendix of the CSR and the code to produce the required graphs associated with the analysis and supporting data files for the CANDA.

The design of the data flow, data handling, population definition, programs, data sets, and output must be agreed upon before initiating serious CANDA development. The programs for the latest NDA were written by one statistician for both pivotal studies to aid in the consistency in the approach. This approach needed a greater lead time to develop highly documented, flexible, generic programs and to test and validate the results. It also lead to a unified environment where the similarity of the programs lead to greater flexibility and interchangeability among the statisticians. In addition, the CANDA system was flexible enough to incorporate ongoing changes made to the statistical methods section in the paper NDA.

Validation Procedures

The data integrity and validation procedures were automatically streamlined by the data extraction procedures stated above. The creation of the CRF tabulations and statistical analyses were done on the IBM 3090 mainframe. This required validation when the data were downloaded from the mainframe. All flat files needed for the CANDA were downloaded from the IBM 3090 mainframe to the workstation through the DECnet protocol on Ethernet and subsequently copied and uploaded back to the mainframe. The mainframe master copy and the workstation copy of the SAS programs, CSR tables, CRF Tabulations and other flat file output were compared by using the ‘word’ processing option in the SupereE utility under SPF (System Productivity Facility) on the IBM. The safety and statistical SAS data files from the mainframe were converted to the binary transport format with the Proc Cport command and downloaded to the workstation and converted back into SAS files using the Proc Cimport command on the workstation. The SAS data files were verified by executing the appropriate program on the workstation, uploading the output and performing the SupereE utility on the mainframe. By comparing the output from the program, not only were the data and analysis results automatically validated, but the version of SAS running on the workstation was also validated.

Validation Outcome

This brings us to the differences found in the mainframe version of the SAS products (version 5.18) versus the PC version of SAS (version 6.03/6.04) versus the workstation (version 6.07).

One inherent difference among these three platforms, is the internal storage of a number and how computational precision is addressed. While this difference is usually of little consequence, it does make a difference when executing statistical methods that count, group, provide univariate information (eg. quantities) or require assessment of a value based on another value (eg. rank transforms). These methods all become inaccurate even with small imprecision, because the exact value is now important. For example if the true value is 3 but the internal representation it 2.9999999999724 then such a value will not be counted with any value of 3 that is correctly represented. It will not be ranked equally with the true “3”, it will not be placed into the correct grouping that may border on a “3”, and will effect the calculation of quantiles.

Rounding decimals on the PC, workstation and mainframe are also different. The PC and the workstation version rounded the decimal number “5” up and the mainframe version rounded the number down. The Proc Univariate on the mainframe defaults the PCITLDEF option to “4” where on the PC the option defaults to “5” also producing a difference if you do not specify the option in your code.

The CANDA System

The Computer Assisted New Drug Application (CANDA) is a comprehensive user friendly interactive
set of tools for use by FDA reviewing staff. The idea that a computer and the computer tools utilized by the company to assist in the assembly of the paper NDA, could also aid the FDA in evaluating the NDA, seems rather obvious.

The challenge is to take these tools that exist, through evolution, in different software packages and hardware platforms, streamline them, improve upon them and integrate them under one hardware platform. This approach not only gives the FDA good tools for their needs but also provides in-house people the benefit of decade old computer processes reanalyzed and enhanced for internal use. This was especially apparent by the creation and implementation of CSR tables, the definition and utilization of patient patient compliance codes, and overall improvements throughout the NDA process.

With this general approach in mind, we formed a CANDA design committee that consisted of a core membership of computer programmer analysts, and revolving membership of medical personnel, statisticians, data coordinators and word processors. The committee knew exactly what procedures worked well and what procedures needed improvements from our first CANDA experience, we knew we had a lot of work to do to meet the FDA statistical requirements section of the system and the computer analysts had a long list of their own ideas they wanted to implement.

Before jumping into the second generation of the CANDA, we took a good long look at the hardware and software on the market. With our very limited staff of two full time programmers, we had to be frugal as to what we could pursue in the limited time we had. We implemented procedures and tools we felt would be long standing building blocks of future enhancements. We did not implement ideas we knew in a short amount of time the market place would address (although sometimes it was hard to resist). Due to our time constraints, we also steered away from procedures and products that were perceived to be difficult or time consuming to integrate with the system as a whole. Both of these aspects will again be analyzed for the next implementation of the system. With this in mind, creation and implementation of the second generation of the CANDA commenced.

As previously stated, there is a great need for early interaction with the FDA. Early in the development cycle, we met with the FDA for their input into the design of the system. This meeting is critical especially if your goal is to have the reviewer utilize your system and not criticize your system. Merck approached the FDA with a proposed CANDA prototype as a basis for discussion. The FDA quickly understood the features and functionalities we had to offer, made suggestions from previous CANDA experience, as well as other suggestions specific to their work environment and the drug under evaluation. All of the ideas that were discussed at that meeting were implemented in the final product. About four months before submission of the CANDA, we met with the FDA again to demonstrate the actual system and gain additional feedback. Additional enhancements suggested at this second meeting were saved for future software updates except for some that could be easily implemented without disturbing the final developmental stages of the system. The second generation CANDA system submitted to the FDA contained approximately four hundred (400) SAS/AF catalog entries which included CBT, program and menu screens, approximately fifty (50) statistical programs, two hundred (200) efficacy data files and one hundred twelve (112) statistical outputs and graphs.

**Key Features and Functionalities**

The FDA was provided with the features and functionalities perceived to be an asset for the reviewer during the evaluation of the NDA. These features (figures 3 and 4) included the following:

**Merck Results:** This option contained two parts, the Safety Data section and the Efficacy Data section. The Safety Data section contained all specified safety data files, CSR tables and CRF Tabulations. The ability to extract the data at the patient level, the record level or all of the data for a given group of variables were available. The ability to identify a group of patients and link to the Ad Hoc Functions option existed.

The Efficacy Data section of the system contained all specified efficacy data files, programs, outputs and graphs. Available functions included the ability to edit a copy of the program, the ability to produce a criteria report to follow selections made during the definition and analyses of a population and the ability to identify a data file and link to the Ad Hoc Functions option.

**Ad Hoc Functions:** There were five major areas under this feature; they were:

- **Data Set/Populations** option provided the ability to select a pre-defined population, create a new population, or select from a previously user defined population. The ability to merge and append data files were also available.
- **Ad Hoc Querying** functions included the ability to create permanent formats to continuous variables, create new variables on data set, include/exclude groups of patients or studies, and conditionally subset the data file. Other features enabled the user to execute a program to count the number of patients and records for each condition placed on the data file and the ability to output a data file and link the file to the other features within the Ad Hoc Functions option.
- **The Statistical Calculations** option provided four functionalities: 1. The ability to create several one way to three way count tables with the option to make one variable unique in the count, 2. The ability to calculate up to six change/percent change calculations identifying the pre and post time variable and value, 3. The ability to create descriptive statistics on numeric variables and output the results to the graph or query option, and 4. The ability to perform two types of correlation statistics. The **Graphics Creation** feature provided for the creation of histograms, line plots with user defined default.
confidence intervals and a confidence calculation, three-dimensional bubble plots and the option to link to SAS/INSIGHT, a three-dimensional graphic software package.

The File Utilities provided the housekeeping functionalities for the system. Included were the ability to copy, delete, rename, display and print files.

Text Summaries: The two Clinical Study Reports and the Overall Summary Report were available within the system. The ability to search on key words and to extract the data from a graph or table were features of the text option.

Reviewers File: This option was a catch-all option which provided a place for FDA special requests to reside. This option included all of the statistical Patient Logs (a listing of all patients efficacy data status by timepoint) and an Investigators Information file that contained a count of patients by protocol, study treatment group and assigned allocation numbers.

Other Software: Other software required by the system or reviewer were available under this option. SAS/INSIGHT was available under the Ulitrix operating system. The ability to create a DOS shell using SoftPC to run DOS programs was available and the DOS version of WordPerfect was included.

SAS: The ability to get to the native software language was available to the reviewer.

Help: There were three types of on-line help: help at the global system level, help at the screen level and help at the field level.

We overhauled and implemented the major components of the ad hoc querying facility, the statistics, graphics, and data set utilities from our first CANDA system. The FDA did not need the ability to browse case report forms within the CANDA. Pop-up windows were used to allow for retrieval of libraries and the selection of data fields on the data set. Pop-up windows were created that displayed the contents of the data file so the user would know how the data was stored in the file to facilitate ad hoc querying.

A SAS macro count program was implemented that counted the number of patients and observations that met each of the subsetting conditions placed on the dataset (table 2). The FDA had stated a need to have a cross-reference between the paper NDA and CANDA. We implemented this in such a manner that the file became as important to users at Merck as it was to the FDA. It would also be one of the few project specific programs within the CANDA that would need modification for future implementations. What the FDA saw in the CANDA as the cross-reference was the NDA volume number, the NDA page number, the NDA table or figure number and a shortened title of the table or figure. What the FDA didn’t see on the screen was the behind-the-scene information the system needed to retrieve the proper data files, programs, outputs, graphs and any other information specific to the reference in the
The Role of the Company Statistician

The statistician plays a major role in the development and success of the CANDA application. The statistician is a link between the objectives of the study and the analyses that are done to meet these objectives. This knowledge is vital during consultations with the CANDA staff and the FDA. Cooperation, creativity, organization and flexibility are four important ingredients needed by the statistician supporting the project. As part of a properly documented analysis, a statistician should be able to help design an analysis system that allows for the tracking, documentation and reproducibility of data and results. The inclusion of a CANDA as part of this system should enhance, not compromise, the statistician's primary objective of performing the proper analysis. In the long run, the CANDA will enable the statistician to spend more time analyzing the data than manipulating and writing analysis programs.

A goal at Merck is to incorporate the CANDA into our in-house data analysis systems so that it is easily exported to the FDA. Future plans for the statistician include migrating their analysis development to the workstation level. The CANDA application will be enhanced to fit a multi-user environment. A user-friendly interface will be created for the statisticians to directly update the paper/electronic NDA cross reference list. They will have the ability to develop and execute their programs within the CANDA system. A library of analysis parameter prompts will be available. Depending on the type of analysis, various subgroups of these prompts will be used prior to program execution to establish the focus of the input data population and the analysis outputs.

Many benefits will result from these changes. Providing the workstation platform to the statisticians will further integrate the CANDA into the normal activities of NDA preparation. The many features and functionalities the system has to offer will be utilized in-house. The statisticians can easily become fluent with the system and will be able to provide earlier feedback to the application developers. In this manner the CANDA will reduce cross system validation and provide a multi-user environment which would promote the sharing of information.

The Future of the CANDA

This generation of the CANDA was a giant leap forward from our first CANDA system. We learned a great deal from our own experience and the experience gained from other companies playing in the CANDA arena. The hardware and software technology is moving so quickly that constant evaluation must occur with each step of development. After a while, the rapid growth of new computer technology might plateau, providing companies the chance to catch their breath and implement a solid working environment for the company that may not be outdated before it is completed. We expect our in-house systems will evolve toward the CANDA we export to the FDA which now has capabilities and features not yet available to our own staff.

Acknowledgments

Many thanks to the following people at Merck for their input and dedication to the CANDA project:

Dan Himmel Joanne Beyer Tom Cook
Margaret Coughlin Cathy Doran Randy Gerard
Mar-yonna Kersten Kim Nosky Sharen Sumter

Special thanks to the following Digital Corporation employees for their technical and sales support throughout the project:

Jeff Janock Greg Emon

Thanks to Tali Axelrod for producing this paper.

Trademarks

Archive Link is a registered trademark of the Archive Link Corporation, Inc.
DEC is a registered trademark of the Digital Equipment Corporation, Inc.
IBM is a registered trademark of the International Business Machines, Inc.
Inquire is a registered trademark of Infodata Systems Corporation, Inc.
QuartetDeck is a registered trademark of the QuarterDeck Corporation, Inc.
SAS is a registered trademark of SAS Institute, Inc.
SPF is a registered trademark of the IBM Corporation, Inc.
WordPerfect is a registered trademark of WordPerfect Corporation, Inc.

Table 2

<table>
<thead>
<tr>
<th>COUNT PROGRAM OUTPUT</th>
<th>NO. OF PATIENTS</th>
<th>NO. OF OBS.</th>
</tr>
</thead>
<tbody>
<tr>
<td>original data set</td>
<td>60</td>
<td>218</td>
</tr>
<tr>
<td>sex eq ‘m’</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>age ge 40</td>
<td>30</td>
<td>130</td>
</tr>
<tr>
<td>(sex eq ‘m’ and age ge 40)</td>
<td>18</td>
<td>118</td>
</tr>
<tr>
<td>serious eq ‘y’</td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td>relation in (possibly ‘probably’ ‘definite’)</td>
<td>19</td>
<td>32</td>
</tr>
<tr>
<td>(serious eq ‘y’ and</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>relation in (possibly ‘probably’ ‘definite’)</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>NUMBER OF PATIENTS AND OBSERVATION PER TREATMENT GROUP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>treatment group</td>
<td>number of patients</td>
<td>number of observations</td>
</tr>
<tr>
<td>Drug A</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Drug B</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>