One aspect of pharmaceutical regulatory affairs is concerned with organizing and auditing clinical project data which are to be used in dossiers submitted to regulatory agencies. When submitting a new drug for registration, information must be compiled on the status of all studies conducted on that project.

For each project, clinical studies are conducted following a prescribed protocol. When complete they are summarized domestically in a standard format, the A to M summary. A to M merely denotes the sections, A, B, C, D, ... M of the summary. The A to M summary can be further classified as either: SOLITARY; having the one study per protocol, INDIVIDUAL; a separate summary for each investigator following a particular protocol, or COMBINED; which includes information from all studies done under the same protocol. All studies, including those cancelled or for which there are no data in house, must be accounted for. The SAS system described in this paper was used as a tool in organizing a large quantity of such multi-source project data.

The particular project under discussion had been underway for a number of years and involved approximately 450 complete or on-going clinical studies. This was a formidable amount of information to organize much less manipulate by hand-log alone. The assignment required a knowledge of the history as well as the current status of the entire project. It was foreseen that many questions would be raised. The answer to each question would require much of the same information to be presented in a slightly different way, each variation requiring many hours of repetitive effort. For example, if the project records were organized in study number order, then there would be questions relating to the information in alphabetical order by investigator; if it were an alphabetical listing, the data probably would be needed by study title or protocol. Keeping half a dozen or more lists, each requiring update as additional information was received, did not seem like an optimal solution. A standard report did not meet the requirements either. Standard reports can be quite cumbersome to work with when only a few pieces of information are needed from each page of a multi-page computer output. The screening and correlation with other non-computerized bits of information, say from hand tabulations, can be tedious, time consuming and error-prone. The situation was complicated by the fact that over the years many departments and many people had worked on the various phases of the project. There was no integrated core of information; although the information was available, only some was computerized.

The first step in solving this problem was to assemble as much relevant information as possible; the second, to decide how to best organize the data much that manipulation would be facilitated.

REQUIREMENTS

Outlining foreseeable requirements was a start. Broken down in system and information requirements, the most important of these seemed to be:

A. System Requirements

1. A simple input mechanism
2. Ability to add additional study information (observations) as new study numbers were assigned
3. No limit on record and variable length
4. On-line access to information
5. Uncomplicated update characteristics
6. Uncomplicated report-generating features
7. Ability to generate an overall summary of all project studies
8. Require, at most, minimal systems and programming support

B. Information Requirements

1. The data set must include all study numbers. Study numbers are allocated sequentially, by project, to signed protocols as they are received.
2. Identify studies

a. which were cancelled or assigned a study number but never started
b. which could be duplicates
c. conducted under particular protocol
d. on which audits of data accuracy have been completed
e. for which additional raw data have been received since the audit
f. included in a prior dossier for registration
g. for which solitary summaries have been written. The study and summary are unique
h. for which individual summaries have been written. Applicable to parallel studies done under the same protocol but written up individually for each investigator

i. included in combined summaries. A combined summary incorporates study information available from the many individual studies done under a particular protocol.

Given the above requirements, SAS was the best choice of the options available, primarily because the whole system could be established and run by someone having a minimal knowledge of computers.

DATA INPUT

During the course of the project, new naming, numbering and study identification conventions were established by the department. The new methods were not, however, retroactive and study information received previously remained in the old format. Allowances had to be made for each convention when assembling this information. Due to the age of the project, the majority of work on the individual studies, summaries, etc., had already been completed. To these, proportionally little would have to be added and updated at a later date.

The information was submitted to batch processing via tape. The tape was generated using the Pertec XL-40 key to disk data entry system. The XL-40 was programmed to prompt for all variables and pre-screen for invalid variable values.

Corrections, additions and deletions to this initial SAS data set were accomplished using the SAS editor procedure under TSO. The "standard" observation was designed for a maximum input of 144 characters. Codes were used in many instances to decrease variable length and speed input.

If a protocol was used for several studies, it was much easier to indicate a protocol number than to include the entire title (especially in the case where one protocol covered more than 40 studies). Combined summaries were also identified by number. Separate dictionaries identifying the codes were created. These were set up so that not only preliminary or working reports, but also detailed reports could be produced without unduly increasing the size of the basic data set. The variables for this SAS data set are identified in Table I.

SUBSEQUENT ADDITIONS AND MODIFICATIONS TO DATA

Once the bulk of the data had been entered and numerous reports generated, additional new reports were requested for which all of the required information was not available in the existing data set.

Rather than reformattting the existing data set, an additional data set was generated. The new data were added using as a key the unique identifier of each observation - the study number. To generate a report incorporating this new information required only a simple merge.

This technique was used to add a country code to the data, as well as in classifying each study by its unit in the New Drug Application. Company defined codes were used instead of country names to speed input and save space. A separate country dictionary was set up which included all countries in which clinical trials might possibly be conducted.

REPORT GENERATION

Depending on the intended purpose, various levels of program complexity were required to produce the final report. In many cases a simple sort and print was all that was required to answer the posed question. Output was generated on-line using a DECKWRITER II or routed to the IBM 3800 laser printer. Examples of reports available under this system include those arranged and printed:

1. by protocol, in study or investigator order
2. by country in which the study was conducted
3. by type or summary written
   a. individual or solitary A-M
   b. combined A-M
4. to list all pre-marketing studies
5. to list all studies previously included in a registration dossier
6. to list all studies with no data in house
7. to list all cases where receipt of new data required a re-audit of a study
8. to show to which unit of the New Drug Application a particular study was assigned

A pictorial summary of the system can be seen in Figure 1.

CONCLUSION

SAS is versatile. It can be applied both to large-scale projects as well as small. It is excellent for organizing data for very specific applications by individuals with little programming background. SAS is outstanding in providing the advantages of computerization to the small scale user.

ACKNOWLEDGEMENTS

The author wishes to thank Frank J. Potter for his technical assistance in setting up this system.
### TABLE I

**DATA SET VARIABLES**

<table>
<thead>
<tr>
<th>NAME</th>
<th>TYPE</th>
<th>DEFINITION/PURPOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>STUDY</td>
<td>NUMERIC</td>
<td>study number</td>
</tr>
<tr>
<td>ST</td>
<td>NUMERIC</td>
<td>identifies extensions to the study</td>
</tr>
<tr>
<td>INVEST</td>
<td>ALPHABETIC</td>
<td>investigators name</td>
</tr>
<tr>
<td>PROTOCOL</td>
<td>NUMERIC</td>
<td>identifies the study by protocol number</td>
</tr>
<tr>
<td>TITLE</td>
<td>ALPHABETIC</td>
<td>identifies the study by an abbreviated title if the study was not assigned a protocol number</td>
</tr>
<tr>
<td>STATUS</td>
<td>ALPHABETIC</td>
<td>identifies studies which are CANCELLED, have NO PROTOCOL or are DUPLICATE</td>
</tr>
<tr>
<td>AM1</td>
<td>ALPHABETIC</td>
<td>(S or T) indicates studies having a SOLITARY (= S) or INDIVIDUAL (= T) A-M summary</td>
</tr>
<tr>
<td>AM2 and AM3</td>
<td>ALPHABETIC</td>
<td>(C1-99) indicates the particular combined summary or summaries in which a study has been included</td>
</tr>
<tr>
<td>COMBINED</td>
<td>ALPHABETIC</td>
<td>identifies the study pairs for which combined summaries have been written. This was prevalent in the older studies where one investigator could have a different study number for each extension but the summary was written to include all of the material.</td>
</tr>
<tr>
<td>CRF</td>
<td>NUMERIC</td>
<td>identifies the studies having unaudited case report forms by giving the number of forms not yet audited</td>
</tr>
<tr>
<td>CER</td>
<td>ALPHABETIC</td>
<td>(CER, CER2-9, PMRT, or A) identifies studies included in the various various clinical experience reports/ dossiers. PMRT identifies pre-marketing studies and A identifies material that has been audited but not yet included in a release.</td>
</tr>
<tr>
<td>COMMENTS</td>
<td>ALPHABETIC</td>
<td>gives pertinent information not covered elsewhere</td>
</tr>
</tbody>
</table>