A TUMOR REGISTRY INFORMATION SYSTEM APPLICATION

William Ingram, University of Florida

INTRODUCTION

The Division of Medical Systems has developed a computerized Tumor Registry Information System (TRIS) for the Shands Teaching Hospital, a 450 bed acute care facility associated with the University of Florida College of Medicine. This registry is designed to provide high-quality, standardized data, for the analysis of cancer treatment and diagnoses, that will (1) ensure that cancer patients receive annual physical examinations, (2) provide follow up data on the current status of each patients' disease and subsequent treatment, and (3) aid the faculty in performing research projects. The tumor registry maintains information on approximately 14,500 patients adding about 2,500 annually.

At the beginning of 1978, the Tumor Registry consisted of 1) two sets of manual card files, a patient file containing data alive and deceased, and a follow up file containing information on patients requiring follow up, and 2) an automated system. The automated system, developed in COBOL in the early 1970's, was virtually duplicated by the manual file. The staff of the Tumor Registry, having discovered several errors in the manner in which the automated system processed data, had little confidence in it and was relying exclusively on manual systems. At this time the volume of the existing files and the rate of information flowing into it had grown to such proportions that the manual registry could no longer keep up with the need to provide detailed case analysis, and physician demand for statistics. Consequently a system improvement project team was created to analyze the requirements to re-establish a computerized registry.

WHERE TO START

After an initial systems analysis the following points were identified as crucial to the projects success: (1) The computerized system had to provide some immediate relief to enable the tumor registry staff to dig out from under their backlog. (2) Because of the need for a quick start, the system had to be able to begin with a limited amount of patient information and to add data fields as staff time permitted. (3) Coding schemas for cancer information were in a state of flux (eg. a new HICDA coding scheme was being introduced, and SNOP was being replaced by SNOMED). Thus, the system must be flexible in allowing edits and updates at a later time. (4) The data from the previous automated system must be converted. (5) As a result of the above points, input forms and their formats were likely to change. (6) Not only were standardized reports required (eg. alphabetic listing of all the cases in tumor registry) but also many varied Ad Hoc reports and statistical data summaries would be required to satisfy the research needs of the physicians.

These needs suggested a computerized system developed within the framework of a software vehicle that provided: (1) A high degree of data independence, which would allow: rapid changes in the input format without affecting the application programs, and permit application program development without concern for the physical data structure. (2) A full set of data manipulation and management features that would allow sorting, selecting, editing, and updating of observations in a wide variety of applications. (3) Report generating capabilities that include both "quickie" reports with formats determined automatically by the system, and "customized" reports with the formats determined by the application program. (4) A method of representing the stored data that was simple and easy to understand.

The TRIS's needs for flexibility and a quick startup precluded development in a higher level language, as the development of needed utility modules (eg. data manipulation, summary statistics, printer plots, report generator, etc.) would be too time consuming. SAS was selected since it provided: the desired programming capabilities, the needed data management facilities, and satisfied many of the desired criteria specified above.[1,2,3] SAS provided an added benefit in its ability to be used by non programming personnel to create special one-time reports.[4]

The first step of the systems analysis was to determine the information reporting needs of the system. From these general needs specific report elements were identified. These data elements were then organized to create the data base and specific report formats were settled on. Once the report formats had been "finalized", the data elements selected, and the initial design of the SAS data sets was accomplished, the input forms were designed. Three factors influencing the design of the forms were the ease of: (1) data entry for the Tumor Registry personnel, (2) keypunching from the data form, (3) updating and adding the new data to the SAS data sets, and (4) using the forms as a temporary manual data file until the data was confirmed as correctly added to the SAS data set.

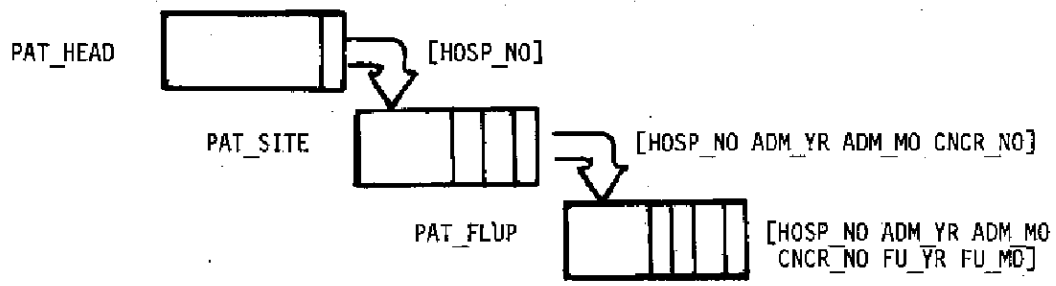

Fig. 1. The SAS data sets used to implement the relations for the TRIS. Relation names are indicated to the left, ruling parts within [].

THE DATA BASE

The data base was organized employing the normalization procedures of relational data base techniques. [1,5,6] The data base was identified as consisting of three relations: (1) Patient Header, (2) Patient Site, and (3) Patient Site Followup. (See Fig. 1)

The patient header consisted of all data items unique to the patient: hospital number, patient name, birthdate, demographic information, date of death, etc. The ruling part, or unique key, is the hospital Number. The Patient Site consists of the data items that are unique to the admission of a patient for a new cancer primary, including: the hospital number, admission date, primary number, descriptive information on the cancer type and treatment information. The ruling part of this relation includes: the hospital number, to link it with the Patient header, the admission date, and the primary number, to identify multiple primaries diagnosed at a single admission. The Patient Site Followup consists of the data elements specific to the follow up of a particular patient's cancer. The Patient Site separate, distinct cancer is followed annually.

Through the use of the SAS MERGE, SET, subsetting IF, BY, and IN comands, these relations may be manipulated to select and combine certain desired observations from the different relation. These observations then form new relations that can be input to SAS's statistical, graphic or report generating procedures.

THE INPUT PROCEDURE

Once the report formats were designed and the structure of the data base defined, the input formats were created. The collection and abstraction of data for input to the Tumor Registry System is a long and involved process. Most of the data to be input must be extracted manually from the patient's permanent medical record. In some cases the summary of initial treatment may not be finalized until several weeks or months after the patient's admission. This aspect, coupled with the fact that the TRIS was to operate in a batch mode, created the need for an intelligent input procedure. (See Fig. 2) Since the input was scheduled to be keypunched monthly, the turnaround time between input document preparation and the running of the data input procedure varied from one to three weeks.

The SAS code for input involved a great deal of syntactic edit checking, including: range checking, valid value checking, and proper coding for ruling part fields. The errors are divided into two classes: field and record errors. The design philosophy was such that: Any input record with a valid ruling part would have the "good" data posted to the data base, the field error flag would be set, and the "bad" data would be signaled in the error report. Conversely, any record with an invalid ruling part would set the record error flag and would only appear on the error report. It was felt that this would provide the least amount of redundant data entry. When updating or correcting data only the proper fields of the ruling part and the desired update fields need be entered.

The SAS data sets are stored on tape as a generation data group. The standard procedure at our installation is to allocate a fixed pool of tapes and rotate the tapes each time a new generation is required. The oldest generation is then reassigned as the newest generation. The details on how to accomplish this are explained elsewhere. [7]

The Semantic checking of the data base is performed on a monthly basis. (See Fig. 2) This involves the comparing of fields to see if they "make sense". For example, checking to see if: the sex and tumor site are valid,

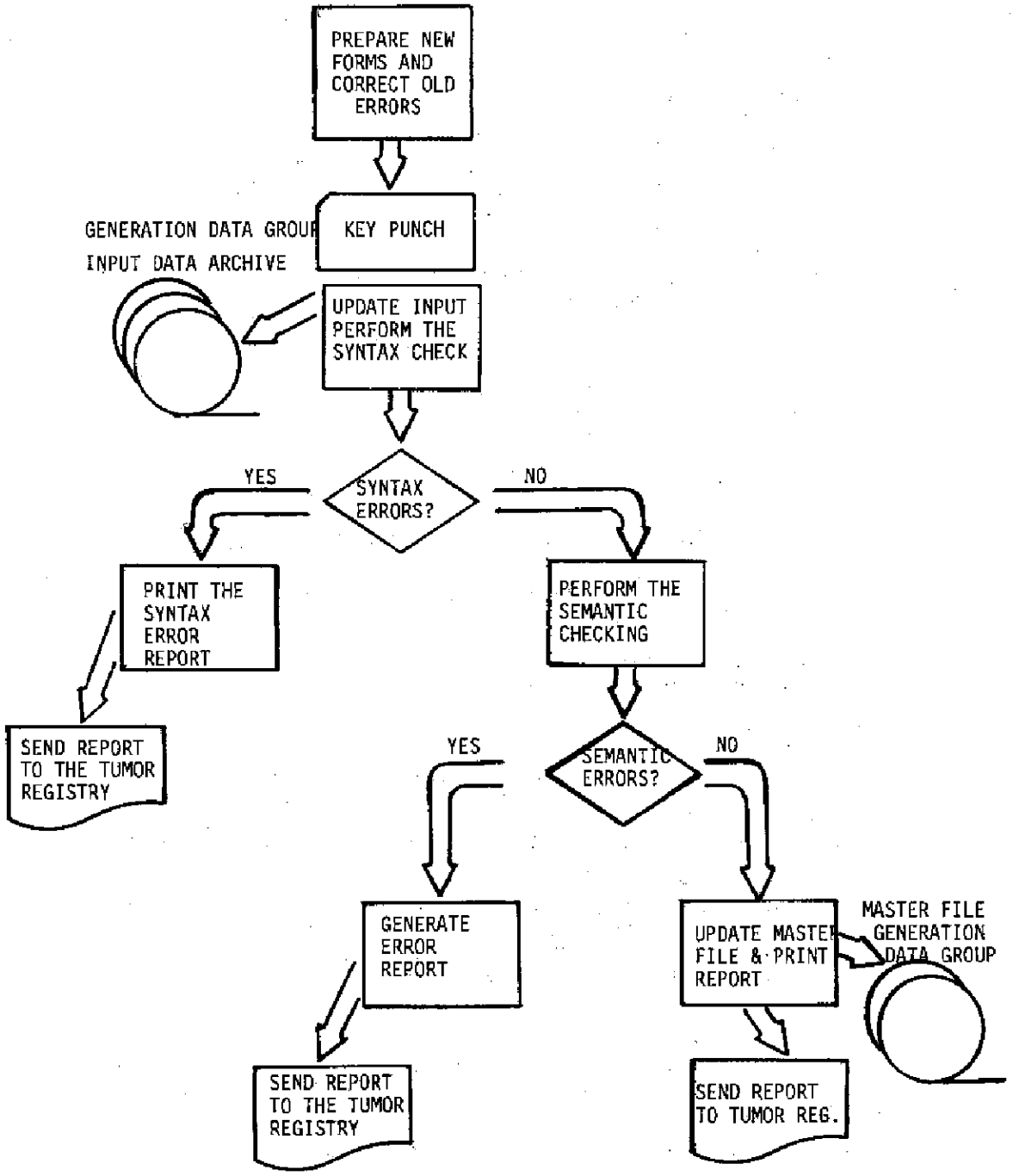

Fig. 2. Monthly processing of data to update the TRIS.

the status on the last follow up correspond to the deceased information, all necessary dependent fields are filled in, etc. In this way, we can automatically audit the data for correctness of form. In addition on an annual basis, a randomly selected number of cases are audited against their charts for correctness. In this manner the high quality of the data is maintained with a relatively low level of personnel involvement.

THE REPORT FORMATS

The types of reports required by the TRIS fall into three categories:

- indexes of the database including,
  - . all living cases in alphabetic order, or hospital number order.

```
//SAS.SYSIN DD *
OPTIONS MISSING=' ';
[ INCLUDE PROC FORMAT ]
MACRO RPT_BODY
  IF FIRST.PAT_NAME THEN
    PUT (PAT_NAME HOSP_NO T_REG_NO AGE SEX RACE AUTPSY ATP_MO ATP_YR)
        (@1 $CHAR18. +1 Z6. +1 Z6. +1 Z2. +1 SEX2. RACE2.
         @125 $1. +1 Z2. +1 Z2.)@;

  IF FIRST.ADM_MO THEN
   PUT (CNCR_NO ADM_MO ADM_YR SITE HISTOLGY DIAG_MO DIAG_YR EXTENT STAGE
        T N PRI_TRT TRT_SUMI)
       (@41 1. +1 Z2. +1 Z2. +1 Z5.1 +1 Z4. +1 Z2. +1 Z2. +1 EXTNT4.
        STAGE3. 1. +1 1. +1 PRTRT4. TSUM4.)@;

  IF FIRST.CNCR_NO  & ^FIRST.ADM_MO THEN
    PUT (CNCR_NO SITE HISTOLGY DIAG_MO DIAG_YR EXTENT STAGE T N
         PRI_TRT TRT_SUMI)
        (@41 1.  +7 Z5.1 +1 Z4. +1 Z2. +1 Z2. +1 EXTNT4.
         STAGE3. 1. +1 1. +1 PRTRT4. TSUM4.)@;

  IF FU_MO^=. THEN
    PUT(FU_MO FU_YR SOURCE STATUS TRT_FAIL TF_MO TF_YR TRT_SUMF TRT_MO
        TRT_YR)
       (@85 Z2. +1 Z2. +1 SRCE5. STAT4. TFAIL5. Z2.
        +1 Z2. +1 TSUM4. Z2. +1 Z2.);

  IF LAST.PAT_NAME THEN PUT;
  IF LAST.PAT_NAME THEN PUT 131*'-';
%
MACRO RPT_HDR
  PUT @1 131*'-' //;
  PUT @1 '|>>>> PATIENT HEADER DATA' @36 '<<<<|'  @41 '|>>>> SITE DATA'
      @81 '<<<<|>>>> FOLLOWUP DATA' @117 '<<<|> AUTOPSY'<';
  PUT @1 131*'-';
  PUT @1 'NAME' @20 'HOSP.  TUMOR  A  S R C ADMIT SITE  HIST DIAGN EXT S  T'
      ' N PRI TRT FOLL  SRCE S   TRMT TRMT  TRT TRMT'
      @125 'A ATTP'/
      @34 'G  E A A MO YR' @60 'MO YR' @70 'T' @77 'TRT SUM MO YR'
      @96 'T   FAIL MO YR SUM MO YR' @125 'U MO YR'/
      @34 'E  X C N' @70 'G' @96 'A' @125 'T'/
      @39 'E C' @70 'E' @96 'T' @125 'P';
  PUT @1 131*'-';
%
DATA HDR;SET TP.PAT_HEAD;
DATA HDR;MERGE HDR TP.PAT_SITE;BY HOSP_NO;
DATA HDR;MERGE HDR  TP.PAT_FLUP;BY HOSP_NO ADM_YR ADM_MO CNCR_NO;
PROC SORT DATA=HDR;BY HOSP_NO PAT_NAME ADM_YR ADM_MO CNCR_NO FU_YR FU_MO;

DATA _NULL_;SET HDR;BY HOSP_NO PAT_NAME ADM_YR ADM_MO CNCR_NO;
  FILE FICHE1 PRINT HEADER=PAGE_HDR;

RPT_BODY

RETURN;

PAGE_HDR:;
  TITLE1  SHANDS TEACHING HOSPITAL & J HILLIS MILLER HEALTH CENTER;
  TITLE2  TUMOR REGISTRY;
  TITLE4  REGISTERED PATIENTS  ---  ALL;
  TITLE5 HOSPITAL NUMBER ORDER;
  RPT_HDR
RETURN;
 *
```

Figure 3

. all cases in tumor site code order
- survival statistics for selected tumor types,
- one time surveys of the data contained in the TRIS.

The first two report types can be developed a priori and need customized output. The last type, the Ad Hoc report, changes each time it is requested and usually will employ the report generator, PROC PRINT, or other SAS determined output (eg. PROC MEANS, PROC FREQ, etc.).

The reports that generate the various required indexes were standardized in the following manner. After analysis of the situation, it was discovered that the major difference between the reports was the sorted order of data and the data elements to be used as control breaks. Thus, we could standardize the body of the report and place it in a MACRO. A separate report header was created for each report and placed in a MACRO. Then the data set manipulation statements were created. In this way the SAS code that generates the report was less than 10 lines. (See Figure 3).

After the reports were designed and implemented, our computer facility acquired a microfiche unit (COM). The conversion of the reports to COM was extremely simple, requiring changing of the FILE statement and adding the appropriate DD card to the JCL. These reports are run monthly following the addition of new data to master files.

The format of the survival statistics reports were also able to be predesigned and included in MACRO's. In addition the selection criteria for classes of patients to be analyzed for survival data was implemented through SAS program statements and is extremely flexible.

The creation of the Ad Hoc reports was solved in part by training the Tumor Registry staff to form their own SAS programs. The most useful tool in this process was the creation of "skeleton", or model programs that the staff could use as patterns for their programs.

WHERE ARE WE?

In the past six months, we have been able to prepare the SAS programs to accomplish our initial goals. The tumor registry is presently staying current with their data input load. We have converted all of the old system data and performed syntatic and semantic checking on the data stored in the TRIS. This checking uncovered a large amount of data correction and "cleanup" work that will consume much of our time over the next four to six months. This audit procedure is extremely important in maintaining confidence in the database.

WHAT NEXT?

As certain functions stabilize in their definition, we will examine the feasibility of implementing them as SAS procedures. For example the survival enhancements. The majority of these represent either features that were never automated previously (eg. automatic printing of followup letters and cards) or the addition of more variables to improve the comprehensive nature of the TRIS data base. I view these developments as healthy and indicative of the success enjoyed by the TRIS project. A large part of the credit for that success must be attributed to SAS for providing the software framework that met our needs.

REFERENCES

1) Ingram, William, In Press, Implementing Relational Database Management Techniques in SAS, Proc. 4th Ann. SUGI.

2) Barr, A.J., Goodnight, J.H., Sall, J.P., Helwig, J.T., 1976. A Users

3) Helwig, J.T. (Ed.) 1977. SAS Supplemental Library Users Guide. SAS Institute. 171 pp.

4) Helwig, J.T. 1978. SAS Introductory Guide, SAS Institute Inc., 83 pp.

5) Martin, James, 1975, Computer Database Organization. Prentice-Hall, 556 pp.

6) Wiederhold, G., 1977, Database Design. McGraw-Hill, Inc., 658 pp.

7) Curry, W., Lezotte, D., Ingram, W., In Press, SAS Used for Data Management and Mathematical Analysis of a Flow Microflourometry