Introduction

The annual survey of its members by the American Hospital Association represents one of the largest and most widely used sources of data on U.S. hospitals. Indeed, for many of the largest and most widely used sources of data on U.S. hospitals. Indeed, for many major research project working with the AHA data.

This paper describes the experience of one major research project working with the AHA data in the hope that such a description will contribute to a better general understanding of this important data set. Specifically, the paper describes the activities which were required to incorporate twelve years of AHA data into the data base of the National Hospital Rate-Setting Study (NHRSS), a four year study being conducted by Abt Associates for the Health Care Financing Administration on the impact of prospective reimbursement legislation on hospitals.

In general, the difficulties encountered in using the AHA data tapes fall into two distinct categories:

- problems related to the management and comparability of the annual surveys—i.e. activities necessary to put the data in a format which is useful for a particular analysis; and
- problems associated with the validity of specific values reported by the AHA—specifically with identifying and correcting inaccurate or missing values for data elements.

This paper discusses these two types of problems and suggests some implications for future uses of these AHA data.

The first major problem in the construction of a longitudinal file from AHA survey data was that over the period of interest (1969-1980), the questionnaire formats and data tape layouts changed annually. In addition, the names and definitions of data elements changed over time, while the codes used for categorical variables sometimes also changed. The first step in the file construction process was to create 12 separate SAS files, one for each survey year, with a common subset of approximately 100 data items and common category codes. These were then interleaved by hospital identifier, creating a master time series file with over 50,000 hospital-year observations. It should be noted that there are a variable number of observations per hospital; since hospitals opened, closed, and merged over the period, there can be from 1 to 12 years of data present for each unique identifier.

The decision to use SAS for this first step was obviously prompted by the 'ragged' format of the input data. This decision did, however, have one significant implication for the remaining file construction and data cleaning procedures. All these steps had to be done using SAS, to avoid the overhead of writing out this large data file into another format. (The master file contains approximately ten million characters.) The project staff developed a number of SAS programs to accomplish the required steps, with the goal of minimizing the personnel and system resources required. The rest of this paper describes those steps, and gives some examples of the SAS code used to carry them out.

Sample Selection

The cost of cleaning and analyzing the entire time series file (comprising all 8,914 hospitals that existed for at least one year from 1969 to 1980) would be prohibitive. Consequently, a 25 percent random sample of acute care, nonfederal hospitals was drawn. The definition of the sampling frame required the construction of three variables: the median length of stay (LOS50) for each hospital over the years present in the master file, and the modal CONTROL (hospital ownership) and SERVICE (type of care provided).

Hospitals vary in the data they report for various reasons. The hospital may undergo a change of ownership or approach. In addition, the responses to some items require subjective judgments by hospital officials, and different hospitals or different responders within a hospital over time may respond differently. For example, the type of service offered may change, or be perceived differently by different individuals. Moreover, some hospitals with facilities for both acute care and long-term care may include the long-term care operation one year and exclude it the next.

Thus in order to characterize hospitals for sampling purposes, it was necessary to use measures of central tendency. Figure 1 shows how length of stay was measured in the ratio of total inpatient days to admissions) and the median calculated in the same step. In general, where one DATA step could be made to do the work of a DATA step followed by a PROC, the former was preferred as more cost effective. Thus Figure 2 illustrates how the modal SERVICE and CONTROL values were generated, using the features available within SAS for programming.
records, the quest for efficiency becomes understandable.

DATA OUT ANALYSIS (KEEP=STATE HOSP LOSSO):
SET IN ALLHOSP (KEEP=STATE HOSP IPDTOT ADMTOT):
BY STATE HOSP:
**** ACCUMULATE UP TO 11 YEARS DATA ****:
ARRAY A(I) XI-XII;
ARRAY B(I) X2-XI2;
ARRAV A(I) XI-XII;
IF LAST.HOSP:
IF N>2 THEN DO:
IF LOS THEN DO:
IF ADMTOT THEN LOS=IPDTOT/ADMTOT;
IF LOS THEN DO:
I+1;
N=N+(I-0-1);
END:
IF LAST.HOSP:
END:
DO OVER C;
END:
END;
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
DO OVER C;
AND:
END:
END:
END:
END:
END:
END:
DO OVER C:
GO TO NEXT;
END;
GO TO LAST;
X=CONTROL:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
DO OVER C;
DO OVER C:
GO TO NEXT;
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
END:
Note that for approximately two percent of the cases, responses on utilization apply to a period other than one year, with days covered (DCOV) ranging from 30 to 720 days. Days covered is computed as the ending date for the reporting period (RPTD) less the start date (RFROM), expressed in internal SAS date format. Percentage changes from year to year can only be calculated on "flow" variables (such as inpatient days or admissions) if they have been first weighted by 365/DCOV. A change of 100 percent upward followed by 50 percent down (or vice-versa) is considered suspect.

Following the temporal edits, two ratios (LOS or length of stay, and OCC or occupancy rate) are computed using the three input variables. Since the ratios share a common element, they can be used to determine which of the three items is in error. If OCC looks alright but LOS is out of range, then the problem must be with admissions; if OCC is invalid but LOS is okay, then the value for beds is suspect. However, if both ratios are outside of expected limits, then the common variable (IPDTOT) is probably in error. Note that both lower and upper bounds are used for each ratio. This is to ensure that the distributions are "trimmed" symmetrically, and not biased upwards or downwards by the data cleaning. The cutoffs are chosen from a frequency distribution of the ratios, as well as a priori based on expected values, in order to ensure that the upper and lower extremes are of approximately equal size. In general, for these data the areas eliminated summed to less than two percent of the distributions.

**Imputation of Missing Values**

Much more prevalent than erroneous values is the analytically comparable occurrence of missing values. In general, the problem of missing values is actually a problem of missing records. Over the period 1969-1978 about seven percent of hospitals failed each year to complete the AHA survey. This problem is also widely dispersed across hospitals, although again more prevalent in smaller hospitals. Missing data on returned surveys frequently can be computed algebraically from other data in the record. For example, a hospital may report long-term admissions and short-term admissions but not the total of the two. Aside from this type of omission, less than two percent of all records contain missing values. These missing values are most likely to be found in those variables which require a special computation, or disaggregation, on the part of the hospital, or which are related to the hospital's fiscal operations. For example, a hospital is less likely to calculate statistical beds or to provide separate counts for full-time, part-time, and total staff for its own planning. Also, it might be noted that an increasing number of hospitals are no longer providing expenditure information. (In 1977, the figure was nearly four percent compared with about two percent three years earlier.) A percentage of data elements will also have been set to missing by the data cleaning process described above. In addition, "lacunas" were sometimes created by the adjustment of fiscal years to a July 1 start date.

As noted above, a number of hospitals reported data for a period other than one year. Potentially more significant, especially for integrating AHA data with data from other sources, is the variation in fiscal years among hospitals. One hospital may be reporting data for July-June, and another for January-December. As the same survey year. The decision was made to use July 1 as the standard start date, that is, any fiscal years ending prior to June 30 would be FY 1977, while March 1, 1976 to February 28, 1977 would be FY 1976. Where gaps were present in the data, dummy records were inserted, so that for each hospital there was a continuous series from the first year of its occurrence in the master file to the last.

Missing data, whether due to nonresponse or data cleaning, were imputed by time series estimation. Figure 4 shows an example of the estimation program. Five procedures were used:

- If adjacent data points exist, take the average;
- If two prior data points exist, extrapolate forward;
- If two following points exist, extrapolate backward;
- If an immediately prior data point is followed by two missing years, interpolate using the prior point and the next valid one; or
- If two missing values are followed by a valid item, interpolate using that value and the last prior one.

The rules were used in the order listed. Note that estimated values are never used to construct others; only data that were originally present are used for imputation.

Following this step, the data were reviewed once again for out of range values and extreme percentage changes, and were then ready for analysis.
FIGURE 4
ESTIMATION PROCEDURE
M~CRO VARS YEAR /* DATA ITEMS TO FILL */
SURPS VEN YOT ADJOT ADJADON SOBOT
STAT80 ADJMP INQDOT PMYDO PAYRES PAYDO
PAYT0T NPAY0EN NPAYFEE NPAYTOT EXPDET
FTE FTHO FTRN FLPN FTH0T FTRD 
MACRO INSVAR 29 % /* COUNT OF VARS */
MACRO INDX1 YEAR-68 % MACRO INDX2 YEAR+68 %
DATA XXX (KEEP=STATE HOSP VARS):
SET IN.AHAEoiT (KEEP=STATE HOSP VARS);
BY STATE HOSP:
IF FIRST.HOSP THEN SlARTzINDX1;
***** INDEX I RANGES OVER VARIABLE LIST *****;
ARRAY X (I) VARS:
ARRAY YR69 (I) A1-A29 ARRAY YR70 (I) B1-B29;
ARRAY YR71 (I) C1-C29 ARRAY YR72 (I) D1-D29;
ARRAY YR73 (I) E1-E29 ARRAY YR74 (I) F1-F29;
ARRAY YR75 (I) G1-G29 ARRAY YR76 (I) H1-H29;
ARRAY YR77 (I) I1-I29 ARRAY YR78 (I) J1-J29;
ARRAY YR79 (I) K1-K29 ARRAY YR99 (I) L1-L29;
RETAIN START A1-A29 B1-B29 C1-C29 01-029 E1-E29
F1-F29 G1-G29 H1-H29 11-129 u1-u29 K1-K29;
LENGTH L1-L29 2;
***** INDEX u RANGES OVER 11 YEARS OF DATA *****;
ARRAY LAST2 (J) rRg9 rRgS YR69 rR70 YR71
VR12 VR73 YR14 YR75 YR76 YR77;
ARRAY LAST (0..1) rRS9 rR6S YR70 YR71 YR72 YR73 YR74 YR75 YR76 YR77;
ARRAY THISYR (0..1) rR69 YR70 YR71 YR72 YR73 YR74 YR75 YR76 YR77;
ARRAY NEXTYR (0..1) rR7Q YR71 YR72 YR73 YR74 YR75 YR76 YR77 YR79;
ARRAY NEXTZ (0..1) YR71 YR7Z YR73 YR74 VR75 YR76 YR77 VR78 VR79 VR99;
U=INDX 1;
DO I = 1 TO NVARS; /* FILL ARRAYS */
THISVR = X;
END;
IF LAST.HOSP;
SlOP"INDX1:
00 u-START TO STOP;
END:
DO I = 1 TO NVARS:
X=THISYR;
IF X < .2 THEN DO:
ELSE IF LASTVR > .2 & NEXTVR > .2 THEN
ELSE IF LASTVR > .2 & NEXT2 > .2 THEN
ELSE IF LASTVR > .2 & NEXT2 > .2 THEN
IF .2 < X < .7 THEN X = .7;
END; END;
VEAR "INOX2;
OUTPUT;
DO u=1 TO 11;
DO I = 1 TO NVARS: /* RESET ARRAYS */
THISVR = X;
END;
END:
DO I=1 TO NVARS: /* SELECT HOSPITAL VEARS */
RETURN;
END;
DATA OUT_AHAEST; MERGE IN.AHAEDIT XXX:
BY STATE HOSP;
Audit and Review
One last set of tests was carried out prior to the final analyses of the data. This was the

testing of all statistical models against a random sample of hospitals. A ten percent
random sample costs about one-tenth as much to run, and allows an exhaustive test of the model.
Many of the analytic models used for the NHAAS are based on log or difference functions;
thus a random sample of observations would not be appropriate. Since these are time series
data, we need to first sample hospitals, then take all the cases for the selected hospitals.
Figure 5 shows how this was done using an indexed access method. This sampling can greatly
is far more efficient than sequentially processing the entire file.

FIGURE 5
RANDOM SAMPLING USING DIRECT ACCESS

THE DATABASE REFERENCED BY THE IN DO STATEMENT HAS 2 MEMBERS:
"FINAL" - 27,543 HOSPITAL-YEAR OBSERVATIONS
"INDEX" - 2673 HOSPITAL RECORDS

DATA XXX (KEEP=); SET IN . INDEX:
RETAIN N 2674 K 100;
/* SELECT 100 HOSPITALS */
/* N = I/OF HOSPS + 1 */
N = N - 1;
IF UNIFORM(O) < K/N;
K = K - 1;
/* K = DESIRED SAMPLE */
/* STOP WHEN K REACHES 0 */
DO I = F'R TO LVR;
OUTPUT;
END:
DATA OUT.SAMPLE; SET XXX; /* SELECT HOSPITAL VEARS */
SET IN. FINAL (KEEP=STATE HOSP VEAR VARLIST)
POINT~I:
As noted at the beginning of this paper, its primary purpose has been to contribute to
a better general understanding of the types of problems associated with using the AHA Annual
Survey. It was also pointed out that the most appropriate response to a specific problem
in any given instance would, in part, depend upon the nature of the analyses for which the data
would be used.

In general, the combination of problems associated with the AHA data base suggests that one
must be careful to avoid using estimation techniques or doing analyses which imply more
precision than the data can support. On the other hand, recognition and proper treatment of
the data problems which do exist can greatly enhance the types of analyses the data will
support.

Frederick Pratter
Abt Associates Inc.
55 Wheeler Street
Cambridge, MA 02138