The Power of the SAS® Date Functions: Creating a Utilization History for Residents of Nursing Homes

Sandra T. Rothwell, National Center for Health Statistics
Mary Ann Bush, National Center for Health Statistics
Ilene Gottfried, National Center for Health Statistics
Dawn M. Scott, National Center for Health Statistics

ABSTRACT

The 1985 National Nursing Home Survey and Followup collected data on a national probability sample of 11,181 persons living in nursing homes in 1984 and 1985. For each person, data included up to 103 admission-discharge date pairs describing stays in nursing homes and hospitals. Dates were not collected in chronological order and were collected from several sources. Thousands of dates were incomplete and hundreds of reported stays conflicted with one another.

The first challenge in organizing this data is making it sensible: imputing missing portions of dates, correcting conflicts, and maintaining “markers” for stays with too little information for imputation.

The second challenge is presenting the data so they can be easily used for varied analyses and still account for variations in nursing home discharge policies.

The task would be almost intolerable without the powerful, easy to use SAS® date functions. In combination with the use of SAS Views®, these functions allow us to build an analytical file that will be valuable for research in long-term care utilization. The work used SAS/BASE® and SAS/STAT® and is applicable to any SAS® system. Intermediate skill level topic.

BACKGROUND:

The 1985 National Nursing Home Survey (NNHS) collected a variety of data about long-term care facilities and their residents. Data were collected on a sample of patients who were current residents at the time of contact with the facility, as well as on a sample of discharges that occurred within the 12-month period prior to the facility contact. To supplement the current and discharged resident components, the 1985 NNHS included a Next-of-Kin (NOK) component to obtain information, not readily available from patient records or nursing home sources, on factors affecting patterns of nursing home and health care facility utilization.

The National Nursing Home Survey Followup (NNHSF) is a longitudinal study that followed the cohort of current residents and discharged residents sampled in the 1985 NNHS. The survey was designed in response to the increasing demand for information on the dynamics of long-term care use. It consists of three waves of data collection conducted from August of 1987 through April of 1990 and provides data on the flow of persons in and out of long-term care facilities and hospitals.

The study was a collaborative project between the National Center for Health Statistics (NCHS), the Centers for Disease Control and Prevention (CDC), and the National Institute on Aging (NIA) of the National Institutes of Health. It was funded primarily by NIA and the Office of the Assistant Secretary for Planning and Evaluation of the Department of Health and Human Services. Figure 1 shows the relationships between the three components of the survey and the three waves of followup. There are 11,181 persons included in the followup study.

THE OBJECTIVE:

One of the purposes of the survey is to study patterns of nursing home and hospital usage. Such topics as the average length of a nursing home stay, the average total number of nursing home days for persons ever entering a nursing home, the patterns of exit and re-entry from nursing homes, the patterns of transfers between nursing homes and between nursing homes and hospitals, and mortality of nursing home residents can all be studied using these five years of survey data.

However, as will be detailed below, the data are extremely difficult to use. These difficulties result from the differences in the six data collection components and from the differences in collection procedures and types of respondents for these components. There is a wealth of information available in the data files that should be explored. Our objective was to combine the five data sets in a way that would make them easier to access and to facilitate a variety of analyses while maintaining the content of the originally released data. We chose to use the SAS system for this project. First, by using the SAS system we could build a data base that could be readily used by any researcher or data processor within NCHS as well as researchers not located in NCHS. Second, most of the difficult work on the data sets involved the manipulation of dates and partial dates. Base SAS software provides a wide array of date functions that, as you will see, made it possible for us to edit and reorganize these data. Finally, since the data would be used for a variety of purposes, we have decided to use the SAS Views as a way of helping the data users select the correct subset of variables and records for different research purposes.

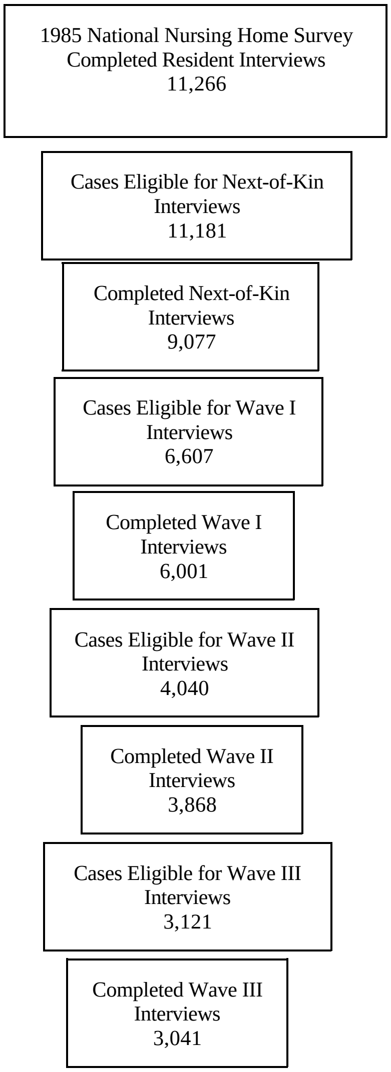
1985 National Nursing Home Survey
Completed Resident Interviews
11,266
Cases Eligible for Next-of-Kin Interviews
11,181
Completed Next-of-Kin Interviews
9,077
Cases Eligible for Wave I Interviews
6,607
Completed Wave I Interviews
6,001
Cases Eligible for Wave II Interviews
4,040
Completed Wave II Interviews
3,868
Cases Eligible for Wave III Interviews
3,121
Completed Wave III Interviews
3,041

Figure 1. 1985 National Nursing Home Survey and Followup

THE DIFFICULTIES:

In the baseline data collection instruments, there were three different interviews. Two of the interviews were conducted in person in the sampled nursing homes. These are the Current Resident Questionnaire (CRQ) and the Discharged Resident Questionnaire (DRQ). Each of these interviews included questions about the admission and discharge dates for multiple stays in the sampled facility as well as admission and discharge dates in other long-term care facilities, that is, not in the sampled home. All the CRQ and DRQ data were collected in the sampled home from an official of the home. These two interviews presented us with a few basic problems.

Information from these two questionnaires is stored on two separate files. It is difficult to combine these files, because the exact same sets of questions were not asked in each interview. The admission and discharge dates were, however, asked in a similar fashion. So, it is possible, though not particularly simple, to combine the files for the purposes of examining dates.

The data collected from the sampled home about the sampled home was fairly complete, though not entirely so. Of the 2,871 records with more than one stay reported in the sampled home , 624 (22%) had date errors requiring correction. A fair number of the dates collected from the sampled homes were missing either portions of the dates (usually the day) or were missing the entire date. There were many cases where either admission or discharge dates were completely unknown. There were also cases where multiple stays were recorded out of chronological order. And there were cases where one stay was either completely imbedded within another stay, or two stays overlapped but did not duplicate each other. Figure 2 illustrates these overlapping and embedding problems. There were also duplicate stays.

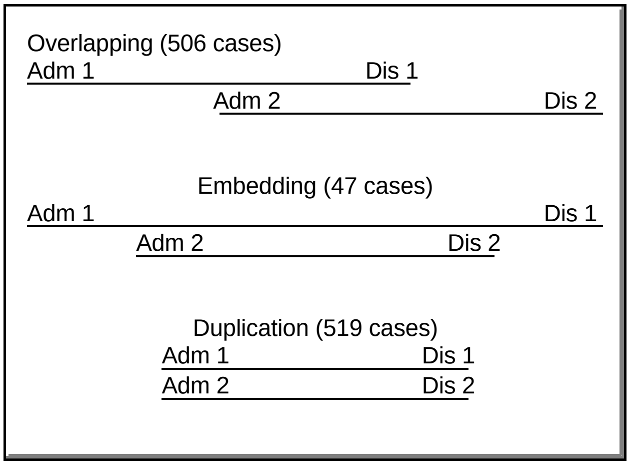

Figure 2. Illustration of overlapping, embedded and duplicated nursing home stays

When the sample home reported information about stays in other facilities, the data were considerably less complete. And these stays were not collected in chronological order at all, but collected so that all the stays in the same facility were grouped together, regardless of order.

The NOK data collection added another layer of problems. The NOK data were collected by telephone rather than in person and the questionnaire collected up to 18 nursing home stays and up to four hospital stays. To begin with, none of the admission and discharge dates (except for the sampled stay) were collected with the day portion of the date. This was because it was felt that most next-of-kin respondents would not recall exact dates but could probably recall the month and year. Of course, there were cases where the month was not known either and cases where even the year was not known. The respondent only reported that the survey subject had been in that nursing home at some time. The NOK respondents also reported hospital stays, with varying degrees of completeness, and only as month and year.

To add to the confusion, the stays reported by the NOK covered part of the same period of time that the CRQ and DRQ questionnaires covered. In many cases the NOK information was in conflict with the CRQ/DRQ information; embedded, overlapping and duplicate stays were not unusual.

Even when NOK dates had no missing parts and were not in conflict with CRQ/DRQ data, it was not possible to impute the day portion of the dates simply by using a value of 15. In many cases, imputing a day of 15, would make an otherwise consistent admission or discharge date, overlap with some existing admission or discharge date from the CRQ/DRQ files.

The followup data collection questionnaires were administered by telephone and sometimes used the same respondent as the CRQ/DRQ files, sometimes the same respondent as the NOK file, and sometimes a different respondent. Fortunately, dates were collected in their entirety, so there was less need to impute days. And the questionnaires did not cover the same time period as the CRQ/DRQ or the NOK, so overlap was less of a problem. However, all the previously described problems still existed to some degree.

In addition we incorporated mortality data from the NCHS National Death Index and Detailed Mortality Files. Any dates imputed or edited from the CRQ/DRQ, NOK, or Followup data had to be consistent with dates of death.

THE TASKS:

The specific tasks of the project were as follows:

1. Clean up all the admission/discharge date pairs. Impute missing data; remove duplicate stays; and correct inconsistent information. There were 6,597 dates with missing data and there were 976 persons with overlapping, duplicated, and/or embedded stays which involved 1,133 dates. For this step, we did not sort the date pairs into completely chronological order, but kept them grouped by facility as they had been collected and sorted in chronological order within these groups.

2. Sort all the stays into chronological order and create variables that would indicate for what facility the stay had been reported. This step would result in a detailed history of long-term care usage for each subject in the survey

3. Combine individual stays which, taken together, describe a complete episode of long-term care. We called this task "redefinition" of the stays and the task is described in more detail below. This step provided a different picture of the detailed utilization history.

4. Identify all those subjects whose first nursing home stay began within 365 days of the CRQ or DRQ interview date. This set of subjects then became a first admission cohort to be used for certain analyses relative only to persons newly entering the long-term care population.

5. Store all the date pairs into a single file along with other survey information such as medical diagnoses, data on functional limitation, and source of payment variables. Provide a mechanism for looking at various subsets of the data. These subsets could include the entire survey cohort, only those survey subjects with a completed NOK questionnaire, only the CRQ subjects, only the DRQ subjects or only the subjects in the first admission cohort.

REDEFINITION:

Not all nursing homes have the same discharge policies. Some nursing homes consider any transfer to a hospital to be a formal discharge and the resident's return is then a formal admission. Others do not consider trips to the hospital as formal discharges, but hold the resident's bed for the duration of the hospital stay. Figure 3 illustrates the two policies. Two subjects, one a resident of Facility A and one a resident of facility B, can have like patterns of usage, but the data for the two subjects will look substantially different. For some studies, the shorter, single stays are the important unit of analysis, for others the longer, combined stays are important. It was necessary to set up the data to handle both needs.

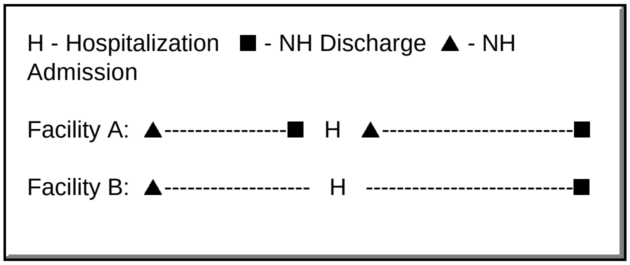
H - Hospitalization ■ - NH Discharge ▲ - NH Admission
Facility A: ▲-----------------■ H ▲--------------------------■
Facility B: ▲------------------- H ---------------------------■

Figure 3. Illustration of the variation in definition of nursing

home stays

THE SOLUTIONS:

1. Cleaning up the admission/discharge date pairs was accomplished by heavy use of many of the date functions. Finding the dates with missing components was easy using the MDY function. Any missing MDY result would mean either missing parts or invalid combinations (such as February 30). If we had to impute dates, we often imputed them halfway between two known dates. The arithmetic properties of dates in the SAS system made this easy to do.

The dates are stored in the MMDDYY format.

```
Obs.     ID   ADM1     DSH1   S1  ADM2     DSH2   S2
1.       AA  061585   079885  A  071885   080985  C

2        BB  071585   081585  A   989898  121585  D
```

Problem 1. Missing day in the first observation.

```
        DATA MISDAY;
             SET ALL;
```

Note: Compute DSH1 halfway between July 1, 1985 and ADM2 (July 18, 1985)

```
  DSHMDY1 = MDY(07,01,85) +
        (( MDY(07,18,85) - MDY(07,01,85))/2); *
   DSH1 = PUT(DSHMDY1,MMDDYY6.);
```

DSH1 = 070985

Problem 2. Missing admission date in the second observation.

```
        DATA MISDATE;
             SET ALL;
```

Note: Compute ADM2 halfway between DSH1 and DSH2.

```
  ADMMDY2 = MDY(08,15,85) +
        ((MDY(12,15,85) - MDY(08,15,85))/2); *

   ADM2 = PUT(ADMMDY2,MMDDYY6.);
```

ADM2 = 101585

* In the examples we show calculations using concrete values, in real work of course, this was all done with substrings and variables.

Figure 4. Uses of MDY function for imputing missing data

Before attempting to correct embedded or overlapping stays, the date pairs describing the stays were sorted within the reporting facilities. After sorting, conflicting stays could be identified by comparing admission and discharge dates. The LAG function was heavily used in this work. Figure 5 includes an example.

Over 23,000 date pairs were examined using these techniques. Out of the 11,181 subjects, 976 subjects were identified with one or more of these problems through Wave I of the followup. Work continues with Wave II and Wave III. Figuring out how to correct these cases was difficult, even using the date functions, but at least finding them was relatively easy.

The following file has been sorted by ID and the dates are stored in the MMDDYY format.

```
Obs.    ID  ADM1   DSH1    S1 ADM2    DSH2    S2

1.       X   071885  080985  C  080785  091585  D
2.       Y   071585  081585  A  100385  121585   D
```

A data set ordered by time was created by using an output statement together with PROC SORT.

```
        DATA ORDER;
          SET ALL;

          KEEP  ID  ADM  DSH  SOURCE;

          IF ADM1 NE ‘ ’ THEN DO;
             ADM  = INPUT(ADM1, MMDDYY6.);
             DSH  =  INPUT(DSH1, MMDDYY6.);
             SOURCE = S1;
                  OUTPUT ORDER;
           END;

          IF ADM2 NE ‘ ’ THEN DO;
             ADM  = INPUT(ADM2, MMDDYY6.);
             DSH  =  INPUT(DSH2, MMDDYY6.);
             SOURCE = S2;
                  OUTPUT ORDER;
             END;
```

Repeat code for up to eight pair or use arrays

```
          PROC SORT DATA = ORDER;
                        BY  ID ADM;
```

The resulting file will have the following configuration allowing for use of the LAG function to check the consistency of the date pairs:

```
Obs.      ID  ADM  DSH     SOURCE

1.        X   9330  9352         C
2.        X   9350  9389         D
3         Y   9327   9358        A
5         Y   9407   9480        D
```

The LAG function code below will show overlapping date problems between the first and second observations:

```
DATA OVERLAP;
```

In addition, 92 cases with a stay of duration 0 days were found. Some of these turned out to be legitimate in that the resident died on the day of admission. The rest were either eliminated or combined with another stay.

2. Once the stay dates were cleaned up within each facility, they were ordered by time without respect to individual facility. Again, the LAG function was used and a calculation of days between discharges and subsequent admissions was used to check on the consistency of the date pairs. When the difference between a discharge and a subsequent admission is less than 0, an overlapping or embedded stay is indicated. Differences of -365 days indicated reporting or data entry errors. The work took several iterations. Facility type was carried along with each date pair. Figure 6 illustrates this work.

With a file sorted as in Figure 5, compare discharge and admission dates for inconsistencies.

```
  DATA REVIEW;
      KEEP ADM DSH DIFF;
      SET ORDER;
   IF (FIRST.ID =0) THEN
       DIFF = ADM - LAG1(DSH);
       IF DIFF LE 0 THEN
          OUTPUT REVIEW;
```

Figure 6 Use of LAG function for final inconsistencies

3. For redefinition, It was also necessary to calculate the number of days between two nursing home stays. We examined the intervals between the sorted nursing home stays and compared them to reported hospital stays determining whether the break between two stays could be attributed entirely to a hospital stay. In some cases the hospital stay data was incomplete. Here, if the nursing home discharge was made to a hospital, we simply subtracted the discharge date from next admission date to see if the interval between the nursing home stays was short (less than 21 days). Of course the two nursing home stays under consideration had to be in the same facility. We packed the data set of sorted stays back into one record per subject and employed arrays to work through each subject’s history, combining two or more stays into single episodes of care and creating array pointers linking these redefined stays to the sorted single stays.

4. Now the selection of the first admission cohort was easy. We simply subtracted the admission date of the earliest stay from the CRQ/DRQ interview date and checked to see if the difference was less than or equal to 365 days. The cases selected became the first admission cohort.

5. We had now developed four sets of arrays describing these long-term care stays. Array one contained the original data. The second array had dates cleaned up, missing parts imputed, and were sorted on admission date within facility. Array three was sorted chronologically without respect to facility and the fourth array held the redefined stays in chronological order. A separate array will hold the hospital data, cleaned up in similar fashion and sorted in chronological order.

   Views of the data set will be created for all subsets generally needed for analysis. In addition to subsetting on types of cohorts, the views can subset on any of the four nursing home stay arrays and can be used to protect sensitive data. Any researcher can request a view for his or her work. All researchers within NCHS will be using the same complicated data set. But the individual view will be simplified for particular purposes.

SUMMARY

As it was, it has taken over two years to clean up and combine the data from the three baseline data sets and from Wave I of the followup. Work is continuing on Wave II and III. In fact, much time had been spent on these files before we decided to do all the work exclusively in the SAS software. The extremely useful definition of dates and the functions that go with this definition as well as the LAG function and arrays, allowed us to work more efficiently and accurately. Rechecking the file after each step became almost automatic. We could calculate the lengths of stays and intervals between stays easily at each step and produce frequencies to look for biases in our work. While the design work and editing decisions had to be done by experienced programming and statistical staff, the re-checking could be done by a junior programmer. As the work progressed, it got easier instead of more difficult. And we expect to create a complete data set: four ways of looking at the data from six files with clearly defined views in an easy to use structure. Currently, this will be the only five year followup of a national sample of nursing home residents. We hope its use will allow us to complete a variety of important studies of long term care utilization.

Sandra T. Rothwell, Mary Ann Bush, Ilene Gottfried, Dawn M. Scott
National Center for Health Statistics

6525 Belcrest Road
Hyattsville, MD 20782

(301) 436-5979
str1@nch07a.em.cdc.gov

References

Gottfried IB, Bush MA, and Madans JH. Plan and Operation: National Nursing Home Survey Followup, 1987, 1988, 1990. National Center for Health Statistics. Vital and Health Stat 1(30). 1993.

Jonas BS, Madans JH, Rothwell ST, Bush MA, Feldman JJ. A method to redefine stays on the 1985 National Nursing Home Survey. Vital Health Stat 2(115). 1992.

National Center for Health Statistics. Public-use data tape documentation: National Nursing Home Survey, 1985. Hyattsville, Maryland: Public Health Service. 1988

National Center for Health Statistics. Public-use data tape documentation: National Nursing Home Survey: Next-of-kin Component. Hyattsville, Maryland: Public Health Service, 1991.

National Center for Health Statistics. Public-use data tape documentation: National Nursing Home Survey Followup: Wave I, 1987. Hyattsville, Maryland: Public Health Service, 1987.

National Center for Health Statistics. Public-use data tape documentation: National Nursing Home Survey Followup: Wave II, 1988. Hyattsville, Maryland: Public Health Service, 1992

SAS Institute Inc. (1990) *SAS® Language Reference, Version 6, First Edition,* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1990) *SAS® Procedures Guide, Version 6, Third Edition,* Cary, NC: SAS Institute Inc.
National Center for Health Statistics. Public-use data tape documentation: National Nursing Home Survey Followup: Wave III, 1990. Hyattsville, Maryland: Public Health Service, 1992.
.