Imputation of Missing Date Information for Injuries and Poisonings
Reported in the National Health Interview Survey

Richard H. Coles (Presenter), Patricia Barnes, Lois A. Fingerhut, Jane F. Gentleman,
Nathaniel Schenker, and Margaret Warner
National Center for Health Statistics,
Hyattsville, Maryland

ABSTRACT

The National Health Interview Survey (NHIS), which is fielded by the National Center for Health Statistics (NCHS), contains questions about medically-attended injury and poisoning episodes experienced by family members during the three months prior to the survey interview date. If the respondent cannot provide complete episode date information (day, month, and year) then the respondent is asked to narrow down the date by identifying the month of injury/poisoning, and, if possible, by stating whether the episode occurred in the beginning, middle, or end of a known month; or by stating how many days, weeks, or months ago the episode occurred. For confidentiality reasons, the date of the interview and the day of the episode cannot be released to the public. Users of the NHIS public use microdata files are therefore provided with only the month and year of the episode, as well as the elapsed time (in days) between the episode and the interview. When the NHIS data are processed at NCHS, the elapsed time is calculated for each episode for which a complete date is known, and if the complete date is not known, imputation is used to select an episode date (or equivalently, an elapsed time) at random that is consistent with the inexact date information provided by the respondent. For example, if the interview date was the 15th of a month, and the respondent reported that the episode occurred in the beginning of the interview month (i.e., between the 1st and the 10th of the month), then the elapsed time must have been between 5 and 14 days, so an elapsed time would be randomly selected from the elapsed time interval [5, 14] using a discrete uniform distribution. In cases where not enough date information was provided even to calculate an elapsed time interval, a type of hot deck imputation is used to select another episode to “donate” an episode date and the corresponding elapsed time. This talk will describe the methods and show samples of the SAS® program used to impute episode dates and elapsed times for NHIS injury and poisoning data.

INTRODUCTION

The NHIS is an annual national household survey of the civilian non-institutionalized population of the United States, conducted throughout the year. The annual NHIS sample now consists of approximately 100,000 persons of all ages who reside in approximately 40,000 households. Trained interviewers from the U.S. Bureau of the Census conduct in-person interviews using computer-assisted personal interviewing. Since its inception in 1957, the NHIS has covered a wide range of health topics, including general health status, acute and chronic conditions, use of health care services, health insurance coverage, and disability and its consequences, as well as basic demographic and socio-economic information. The NHIS questionnaire contains three major sub-modules, which cover the entire family (about whom a knowledgeable adult responds), a randomly-sampled child (about whom a knowledgeable adult responds), and a randomly-sampled adult (who responds for him/herself).

Questions about injuries and poisonings are asked as part of the family sub-module. Information about all injuries/poisonings that were reported to have occurred within the last three months and to have been medically attended is collected by the survey, processed later at NCHS, and then released as part of the annual NHIS public use microdata file. For each such injury/poisoning episode, the interviewer collects as much information as possible about the date of the episode, the cause of the injury/poisoning, where the injury/poisoning occurred, what activity the injured/poisoned person was engaged in, what body part was affected by the injury/poisoning, the type of injury/poisoning, and whether days were consequently
missed at work or school. When the data are processed at NCHS, ICD-9-CM nature of injury and external cause codes are assigned to each injury/poisoning incident.

Episode date information is obtained by the NHIS in two stages. In the first stage of questioning, the respondent is asked whether the episode occurred during the past three months, and a specific date 91 days ago is provided as a reference. If the respondent indicates that the episode occurred within the 91-day window, the respondent is asked in a second stage of questioning to provide a day, month, and year when the episode occurred. If the month is not reported, the respondent is asked to give an elapsed time since the episode occurred, specified in the respondent's choice of unit (days, weeks, or months). If the month but not the day of the month is reported, the respondent is asked whether the episode occurred in the beginning, middle, or end of the month.

If the complete date (day, month, and year) of an episode was not provided or cannot be determined from the information provided by the respondent, then a complete date is imputed during data processing, consistent with whatever information was provided by the respondent. This paper describes the imputation algorithms developed by NCHS and the SAS code used to carry them out.

**SAS DATES**

The SAS System represents dates internally as the number of days since a reference date. The reference date (“date zero”) used for SAS date values is January 1, 1960. Thus, for example, October 17, 1991 is represented by the SAS System as 11,612. SAS date values for dates before January 1, 1960 are negative. Any numeric variable in a SAS data set whose values are represented in this way is called a SAS date variable. Converting two dates to two date variables and then subtracting the earlier date variable from the later date variable yields the length of time between the two dates. SAS calculations account correctly for leap years; for example, subtracting February 28, 2004 from March 1, 2004 results in a difference of two days.

**SAS RANDOM NUMBERS**

To generate a sequence of random numbers the SAS random number generator uses a starting number, or “seed.” Different seeds will yield different sequences of pseudo random numbers, and the same seed will yield the same sequence each time it is used. If the program specifies a seed that is less than or equal to zero (e.g., “aa= RANUNI(0);”) or if no seed is specified, then a seed is set by the SAS routine based on reading the time of day from the computer's clock, in which case a different sequence of random numbers is produced every time the program is run. To provide reproducibility in testing and running the injury/poisoning data processing software, the injury/poisoning data processing program uses a specified seed (e.g., “Seed = 27303436;”).

**METHODS AND EXAMPLES**

NHIS interviews are conducted using laptop computers. In addition to capturing information about the dates of injuries and poisonings, the survey-administration software captures the actual date on which the injury/poisoning questions were asked (which may be sooner than the date on which the entire NHIS interview was completed). It can be assumed that all day-month combinations collected from NHIS respondents are valid (e.g., no February 30) because the survey-administration software checks the validity of the date during the interview. Later, during data processing at NCHS, a complete date (day, month, and year) is ascertained for each injury/poisoning, as well as the elapsed time in days between the date of the injury/poisoning and the date when the injury/poisoning questions were asked. If necessary, imputation is used to obtain an injury/poisoning date and an elapsed time. When either of these quantities is known or has been imputed, the other may be obtained from it by using operations on SAS date variables. The general procedures for determining those two quantities are as follows:

1. If the complete date (day, month, and year) of the episode is known, then no imputation is necessary.
2. If a complete date was not provided, but it is known that the injury/poisoning occurred during the beginning, middle, or end of a known month, then a random date within the part of the month specified is generated.

3. If a valid month and year were provided, but the part of the month is unknown, then a random day within that month is generated.

4. If the elapsed time in days was provided, then the day, month, and year are determined with certainty from that information; no imputation is necessary.

5. If the elapsed time in weeks or months was provided, then the date that was exactly that many weeks or months ago is not used as the episode date; rather, a random episode date is generated that is within about a half week or a half month, respectively, of that exact date.

6. Random dates were generated as described above by assuming a discrete uniform distribution and generating a random episode date within the time frame specified or implied by the respondent. If none of the above conditions (1-5) occurs, or if certain invalid situations are identified, a type of hot deck imputation is used to obtain an episode date. Hot deck imputation is a method of filling data gaps with data from other observations (“donors”) in the sample at hand.

DETERMINING THE ELAPSED TIME FROM A COMPLETE DATE

To determine length of time in days between a valid episode date and the day the injury/poisoning questions were asked, we convert the dates to SAS dates and calculate the difference between them:

```sas
IF ('01' <= Episode_day <= '31') AND ('01' <= Episode_mon <= '12') AND ('2006' <= Episode_year <= '2007') THEN
DO;
Episode_SASD = INPUT(Episode_mon || Episode_day || Episode_year, MMDDYY8.);
Interview_SASD = INPUT(Date_of_Interview, MMDDYY8.);
Length_of_Time = Interview_SASD - Episode_SASD;
/* Note: Error if date of interview before date of episode */
IF Length_of_Time < 0 THEN
Length_of_Time = 0;
END;

DISCRETE UNIFORM IMPUTATION USING THE ELAPSED TIME INTERVAL

For each injury/poisoning episode, an elapsed time interval is calculated from the date information provided by the respondent. The lower bound of the elapsed time interval is the number of days between the latest date when the episode could have occurred, consistent with the reported date information, and the date when the injury/poisoning questions were asked. The upper bound of the elapsed time interval is the number of days between the earliest date when the episode could have occurred, consistent with the reported date information, and the date when the questions were asked. If the episode date is completely known (day, month, year), then the lower bound is equal to the upper bound, and both boundaries are equal to the number of days between the episode date and the day the questions were asked. When the width of the elapsed time interval is greater than zero, then a random elapsed time is selected assuming a discrete uniform distribution over the interval.

It is possible (but unlikely) for one or both of the elapsed time interval boundaries to be negative. Negative boundaries occur when the window of possible episode dates, based on what the respondent reported, extends into the future or is entirely in the future. These situations are clearly the result of reporting or capture errors, so the injury/poisoning data processing program makes a correction when possible.

The algorithm to impute a valid elapsed time from an elapsed time interval, accounting and correcting for the possible occurrence of negative elapsed time interval boundaries, is as follows: If both boundaries of the elapsed time interval are ≥0, and they are not both equal to zero, impute an elapsed time from the
interval using a discrete uniform distribution. If the lower bound of the elapsed time interval is \(<0\) and the upper bound is \(>0\), set the lower bound to zero, and impute an elapsed time from the resulting interval. If the lower bound of the elapsed time interval is \(\leq 0\) and the upper bound is equal to 0, set the lower bound to zero, and set the elapsed time to zero; imputation is not needed.

Hot deck imputation is used when the information reported about the episode date cannot be used to obtain a reasonable elapsed time interval—i.e., for cases for which a day, month, and year were not reported or could not be deduced—and for cases where the possible dates as reported all occur after the date of interview. For each episode undergoing hot deck imputation, another episode (a donor) must be selected to provide the missing date information. In choosing donor episodes, the underlying assumption is that hot deck imputation cases are more similar to discrete uniform imputation cases than they are to cases for which no imputation at all was needed. Hot deck donors are thus selected only from among cases that underwent discrete uniform imputation. Hot deck imputation is therefore necessarily deferred until all discrete uniform imputation is completed.

**PICKING A DATE WITHIN THE REPORTED PART OF THE MONTH**

If the respondent reported that the episode occurred in the beginning, middle or end of the month then we generate a random day number of 1-10, 11-20, or 21-last day of that month, respectively. For example, suppose that the episode was reported to have occurred at the end of September, in which case a day number between 21 and 30 needs to be selected at random.

An example of one method recommended by SAS to pick a number between two known numbers is as follows:

\[
X = \text{CEIL} (\text{RANUNI}(\text{Seed}) \times 10);
\]

picks a number between 1 and 10. Therefore, the statement

\[
X = 20 + \text{CEIL} (\text{RANUNI}(\text{Seed}) \times 10);
\]

provides a number between 21 and 30.

An alternate method, used in our program, is as follows:

\[
\text{LOW\_NBR} = '21'; \\
\text{HIGH\_NBR} = '30'; \\
/* Pick random number within High and Low intervals */ \\
X=\text{ROUND}\left(\text{RANUNI}(\text{seed})\times(\text{HIGH\_NBR} - \text{LOW\_NBR})\right) + \text{LOW\_NBR};
\]

**PICKING A DATE WITHIN THE REPORTED MONTH**

If the month is known, but not the actual day within the month, then we randomly pick a day between 1 and the number of days in that month:

\[
/* Determine number of days in the month for ending interval */ \\
\text{LOW\_NBR} = '01'; \\
\text{IF MONTH IN ('01', '03', '05', '07', '08', '10', '12') THEN} \\
\text{HIGH\_NBR} = '31'; \\
\text{ELSE IF MONTH IN ('04', '06', '09', '11') THEN} \\
\text{HIGH\_NBR} = '30'; \\
\text{ELSE IF MONTH = '02' THEN} \\
/* Note: Leap Year & February */ \\
\text{DO;} \\
\text{IF YEAR IN ('2008', '2012', '2016') THEN} \\
\text{HIGH\_NBR} = '29'; \\
\text{ELSE HIGH\_NBR} = '28'; \\
\text{END;}\]
/* Pick number within High and Low intervals */
X=ROUND(RANUNI(Seed)*(HIGH_NBR - LOW_NBR)) + LOW_NBR;

PICKING A DATE WHEN THE ELAPSED TIME WAS REPORTED IN WEEKS OR MONTHS

For episodes for which the elapsed time was reported in weeks, the elapsed time interval boundaries are set to 7 (reported elapsed time in weeks) ± 3. For episodes for which the elapsed time was reported in months, the elapsed time interval boundaries are set to 30 (reported elapsed time in months) ± 15. Using weeks as an example, the elapsed time interval contains seven possible elapsed times:

X = Nbr_weeks * 7;
LOW_NBR = X - 3;
HIGH_NBR = X + 3;
/* Pick random number within High and Low intervals */
Y=ROUND(RANUNI(Seed)*(HIGH_NBR - LOW_NBR)) + LOW_NBR;

For example, if the episode was reported to have occurred two weeks ago, the elapsed time is between 11 and 17, with an interval width of 7 days.

HOT DECK IMPUTATION

Hot deck donors of date information are selected only from among those cases where imputation was performed assuming a discrete uniform distribution between the lower and upper bounds of the elapsed time interval of width greater than zero. Note that in selecting donors, sampling with replacement is used, so it is possible for the same donor to be used more than once.

In the SAS program, a variable is set that identifies:
- a) which episodes have completely-known dates and do not require imputation;
- b) which episodes undergo imputation using a discrete uniform distribution assumption; and
- c) which episodes undergo hot deck imputation.

More than one pass through the entire set of episodes is required in order to postpone hot deck imputation until after all of the cases requiring discrete uniform imputation have been dealt with. In our program, the file of episodes is separated into three files (“good”, “donor”, and “needs donor”). We place unique sequential identifiers on each donor record (in this example DCASENBR) and place the highest number of the donor record ID into a macro variable (N_DONOR) to be used in the step that immediately follows it.

***************************************************;
* Number the donors (ID), retain largest number *
* Note: only keep fields to donate to recipient *
***************************************************;
Data Donors(KEEP=Field_A Field_B DCASENBR);
  set DonorRecords END=lastobs;
  DCASENBR+1;
  IF lastobs THEN
    CALL SYMPUT ('N_DONOR',PUT(DCASENBR,5.));
Run;

In the preceding step, after the last record is processed “IF lastobs THEN” the highest sequential ID is placed into a macro variable. A RETAIN causes a variable to retain its value from one iteration of the DATA step to the next. In this case, it is not needed because the code is “DCASENBR+1;” not “DCASENBR=DCASENBR+1;”. The macro variable N_DONOR will be used in the next step as the highest number from which to pick an ID number.

The step below picks a number from 1 to the highest donor record ID, the macro variable. This provides each record an identifier to match against the unique identifiers (DCASENBR) on each of the donor
records. It also allows, with equal probability of selection and with replacement, the same donor to be used for multiple records needing donations.

***********************************************************;
* Pick a number from 1 to largest for a matching donor ID *;
* Note: Fields to be replaced are dropped *
***********************************************************;

Data NeedDonor;
    Set HotDeckRecipient(Drop=Field A Field B);
    /* Pick a Donor = random number 1 to maximum donor ID number */
    DCASENBR = INT(round(ranuni(seed)*(&N_DONOR - 1)) + 1);
Run;

Merge the donor records with those needing donors, keeping only the records that received donor variables.

***********************************************************;
* Merge donor time periods into recipient episodes *
* Note: only records needing donor information are kept *
***********************************************************;

PROC SORT DATA=NeedDonor;
    BY DCASENBR;
Run;

Data RecipientUpdated(DROP=DCASENBR);
    MERGE Donors (in=donor_f)
        NeedDonor (in=needs_f);
    BY DCASENBR;
    IF needs_f and donor_f;
Run;

All that remains now is to merge the original „good“ records, the donor records and the recipient records back together into one sorted file. The donor id is dropped.

CONCLUSION

The NHIS questionnaire fielded by the National Center for Health Statistics gathers information about the dates of injury and poisoning episodes. Some of those dates are incompletely specified. In our 2004 NHIS files, for about 75% of the cases, the respondent provided complete episode dates (day, month, and year); for 20% of the cases, the episode was reported to have occurred in the beginning, middle or end of a month; for 3% of the cases, only the month was reported; for 1% of the cases, the elapsed time in days, weeks, or months was reported; and for less than 1% of the cases, the reported date information was so insufficient that hot deck imputation was needed. The steps detailed above were able to handle all of these cases.

ACKNOWLEDGEMENTS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

CONTACT INFORMATION

Please contact the authors with your comments and questions:

Richard H. Coles
National Center for Health Statistics
3311 Toledo Rd. Room 2217
Patricia Barnes
National Center for Health Statistics
3311 Toledo Rd. Room 2318
Hyattsville, MD 20782
Work: (301) 458-4084
Email: PBarnes@cdc.gov

Lois A. Fingerhut, M.A.
Special Assistant for Injury Epidemiology
Office of Analysis and Epidemiology
National Center for Health Statistics
3311 Toledo Rd. Room 6418
Hyattsville, MD 20782
Work: (301) 458-4213
Email: LFingerhut@cdc.gov

Jane F. Gentleman, Ph.D., Director
Division of Health Interview Statistics
National Center for Health Statistics
3311 Toledo Rd. Room 2217
Hyattsville, MD 20782
Work: (301) 458-4233
Email: JGentleman@cdc.gov

Nathaniel Schenker, Ph.D.
Senior Scientist for Research and Methodology
National Center for Health Statistics
3311 Toledo Rd. Room 3209
Hyattsville, MD 20782
Work: (301) 458-4483
Email: NSchenker@cdc.gov

Margaret Warner, Ph.D.
Office of Analysis and Epidemiology
National Center for Health Statistics
3311 Toledo Rd. Room 6424
Hyattsville, MD 20782
Work: (301) 458-4556
Email: MWarner@cdc.gov