AN EVALUATION OF A METHOD USED TO IMPUTE RESIDENCE DATA

D.E.B. Potter and Peter J. Cunningham, Agency for Health Care Policy and Research
D.E.B. Potter, 5600 Fisher's Lane, Room 18-A55, Rockville, Maryland 20857

KEY WORDS: event histories, minimum distance function, nursing homes

Introduction

The analysis of event histories, sometimes referred to as failure-time data, has become wide spread in the fields of statistics, economics, sociology, and epidemiology during the last fifteen years (Petersen, 1991). Researchers use these data to model failure time distributions, i.e., how durations in various social or biomedical states depend on exogenous covariates (for example, Cox's (1975) proportional hazards model). However, event history data, like other sample survey data, are subject to nonresponse and associated nonresponse bias.

This paper builds upon previous work (Potter and Cunningham, 1990) on the imputation of missing event history data. In both analyses, the event history data are residence data from a sample of persons who spent any time in a nursing or personal care home (NH) during 1987. While the reference period is short (only a year), the study population includes a large number of persons with multiple events (in this case stays in nursing homes, hospitals, etc.). Thus, in developing the imputation method it was necessary to take into account the potential for imputing multiple events to persons with incomplete data due to item or partial (wave) nonresponse.

The first part of the paper describes the National Medical Expenditure Survey (NMES) Institutional Population Component (IPC), the data source for this work. This is followed by a brief summary of the methodology used to construct the year long residence (event history) profiles. The analysis characterizes persons with missing residence data and compares them with those with complete data. The paper also includes a brief description of the imputation methods and an evaluation of the effects of imputation on nursing home utilization estimates. Lastly, we provide an assessment of the imputation on the precision of NH utilization and expenditure estimates.

Sample Design, Estimation and Data Collection Methods

Sample Design

The adopted NMES institutional population survey is a stratified, three stage probability design with facility selection in the first two stages. Current residents (residents on January 1, 1987) and admissions (persons admitted between January 1 and December 31, 1987) were sampled within participating facilities at the third stage. Three explicit sampling strata were used to select the facility sample: nursing and personal care homes; facilities certified under the U.S. Health Care Financing Administration's Medicaid program as Intermediate Care Facilities for the Mentally Retarded (ICF-MR) with 3-15 beds; and other facilities for the mentally retarded (Cohen, Flyer and Potter 1987). Interviewers made four distinct visits to each cooperating facility at approximately four month intervals to facilitate sample selection and data collection in the institutions. January 1 residents were sampled at the first visit and admissions were sampled during the subsequent visits (Edwards and Edwards, 1989).

The IPC facility sample consisted of 851 eligible nursing and personal care homes and 730 eligible facilities for the mentally retarded; however, these analyses are limited to the NH sample. The IPC facility level response rate for NH's was 95.2 percent. Of the 810 responders, 99.4 percent (805 facilities) allowed sample selection of current residents, and 93.6 percent allowed both current resident and admission sample selection for all rounds of data collection (758 facilities) (Potter and Cunningham, 1990).

The current resident sample consisted of 3,392 eligible residents in nursing and personal care homes. The admission sample consisted of 2,608 eligible "new" admissions. New admissions were defined as individuals who were admitted to the sampled facility during 1987 and had no prior admissions to that facility during the survey year. When combined, the current resident and new admission samples form a sample of 6,000 nursing home users, i.e., persons who at any time during 1987 used a nursing or personal care home.

Estimation

To meet the strict estimation requirements imposed on all NMES primary survey components, the sample was restricted to only those individuals who had facility use and expenditure data for at least a third of their period of eligibility (Cohen and Potter, 1990a). In NMES, the minimum part-year response requirement followed the approach taken in the 1980 National Medical Care Utilization and Expenditure Survey (National Center for Health Statistics). Consequently, the 246 SP's with less than a third of their facility use and expenditure data were considered complete nonrespondents. An additional 169 sampled persons were actual nonrespondents. Thus, of the sampled eligible population of 6,000 persons, 5,585 (93 percent) are used for NMES Institutional Population Component estimation purposes (Table 1) (Cohen and Potter, 1990b). A standard nonresponse weight adjustment is used to adjust for the potential selection bias associated with the exclusion of the nonrespondents.

Data Collection Methods

Data were collected under the sponsorship of the Agency for Health Care Policy and Research (AHCPR) by Westat and NORC using the methods detailed in Edwards and Edwards (1989).

Sampled persons were followed throughout 1987. For those who left the facilities in which they were selected, data were collected up to the time of discharge. If a sample person (SP) entered another IPC-eligible NH, the institutional data collection procedures were continued in the new facility. There was no retrospective institutional data collection for persons sampled as admissions who experienced a stay in another NH in 1987 prior to their stay in the sampled facility. Rather, this information was collected from next-of-kin respondents in order to obtain complete 1987 reference period data.

Data collected from NH respondents included facility level characteristics, and for sampled persons, data for:

health status, demographics, socio-economic characteristics such as insurance coverage, residence history, facility use and expenditures, and sources of payment.

In order to obtain data for periods of time during 1987 when the sampled person was not a resident of the nursing home, the IPC contained a Survey of Next of Kin (SNK). Survey respondents were persons residing in the community (usually relatives) who were knowledgeable about the sampled person. Up to three SNK interviews, the number depending upon when the SP was sampled, were done by telephone, using a computer-assisted telephone interviewing (CATI) system, from the offices of the data collection contractors. The system allowed for the collection of very complicated residence history information as well as use and expenditure information. For 79 percent of the nursing home users sampled, at least one SNK interview was completed with a community respondent. When no suitable community respondent could be identified, facility staff were asked a subset of the SNK questions.

Constructing Year-Long Residence History Profiles From Multiple Data Sources

Because of the operational difficulties in collecting information from the SNK respondent that was distinguishable from information already obtained from facility respondents, the design of the next-of-kin survey specified the collection of information about the SP for periods of time when the SP was in a NH as well as, for periods when the SP was not in a NH. This resulted in multiple data sources with conflicting data, especially for the period of time the SP was not in the NH. While this facilitated data collection operations, it made the construction of 1987 residence profiles a complex methodological problem. Several hierarchical assumptions were made to construct the profiles. These were detailed in our previous work (Potter and Cunningham, 1990) and are summarized below:

1. NH's could accurately report periods of NH use for persons residing in their facilities.

2. NH's could accurately report discharges from their facility to an acute care hospital and subsequent readmission from the hospital back to the NH.

3. In additional to interim hospital stays (which were previously considered under assumption number 2) NH's could accurately report data on the stay that occurred immediately after discharge from the NH.

4. Excluding all post-discharge stays reported by the facility (previously considered under assumption numbers 1, 2 and 3), next-of-kin respondents could report data for all 1987 stays that occurred after discharge from the sampled facility.

5. Next-of-kin respondents would be better able to provide information on the period of time in 1987 that occurred prior to admission to the sampled NH (this was only applicable for persons sampled as admissions).

6. For persons with no next-of-kin respondent, NH's could report for the period of time in 1987 prior to admission to the sampled NH.

7. Persons were considered to have died in the hospital if discharged from a NH to an acute care hospital and subsequently reported as deceased within three weeks of discharge.

By making these assumptions and piecing together residence data on multiple stays from multiple data sources residence profiles were constructed for at least some of the time during calendar year 1987 for all persons. Graphic representations for a sample of constructed residence profiles are shown in Figures 1-10b.

Characteristics of Persons with Missing Residence Data

Residence profiles were considered complete (the location of the SP was know for everyday in 1987) for 90 percent of the responding population (84 percent of the sampled population). There was some differential by sample type (those sampled as current residence versus those sampled as admissions). For SP's sampled as current residents, residence profile data were considered complete for 97 percent of the responding population (3,100). The admission sample completion rate was 82 percent (1,944). (Table 1).

Table 2 shows the distribution of persons with partial and complete residence data by type of missing residence data. Of the 541 SP's with some missing data (the Partials), approximately 39 percent were missing only post-discharge data (i.e., the residence data associated with the time period after first discharge from the sampled NH). Persons with unknown discharge status from the sampled facility were classified as missing post-discharge data. A little more than half (53 percent) of the partials were missing only pre-admission residence data (i.e., the residence data associated with the period of time prior to first admission to the sampled NH and only applicable to persons sampled as admissions). Only 8 percent were missing both pre-admission and post-discharge data (calculated from Table 2). Again there was a differential by sample type, with approximately 80 percent of the partials being sampled as admissions.

It was hypothesized that characteristics of the Partials might differ depending upon the kind of missing residence data, i.e., that persons missing post-discharge data might be different from persons missing pre-admission data. As a consequence, persons with missing residence data were divided into three subcategories for analytical purposes. Since there appears to be some differential by sample type, the groups created also control for sample type. The following partial groups were created: $partial_1$, persons sampled as current residents and missing only post-discharge data ($n = 109$); $partial_2$, persons sampled as admissions and missing only pre-admission data ($n = 288$); and a residual category, $partial_3$ ($n = 144$), consisting of the remaining persons sampled as admissions and missing post-discharge data ($n = 103$), and those persons missing both post-discharge and pre admissions residence data ($n = 41$).

Table 3 contrasts the population of persons providing complete residence profile data to the three partial groups for 24 selected characteristics. The domains considered

included characteristics of the sampled NH and person-level characteristics for socio-demographic and health status measures. The chi-square test of independence was used to assess statistically significant differences between the selected characteristics and residence data completion status (a complete or partial residence profile). In total, three sets of comparisons were made: (1) persons sampled as current residents with complete or $partial_1$ residence data: (2) persons sampled as admissions with complete or $partial_2$ residence data; and (3) persons sampled as admissions with complete or $partial_3$ residence data. To obtain variance estimates and associated p values the software program SUDAAN (Shah et. al, 1989) was used. This program considered the complex nature of the NMES IPC design by using the Taylor Series linearization method of variance estimation.

Among the current residents, only characteristics associated with the sampled NH location (SMSA status) and certification status -- the latter an indication of whether the facility is certified by the Health Care Financing Administration (HCFA) for reimbursement under the federal Medicare and/or Medicaid programs -- were found to be associated with the presence of complete or partial residence data ($p = .05$). The $Partial_1$s were more likely than their complete counterparts to have been sampled in a NH located in a SMSA and three times as likely as their complete counterparts to have been sampled in a non-certified facility. The latter is not surprising given the previous finding that total facility nonresponse for NH's in the NMES IPC is associated with an absence of Medicare/Medicaid certification (Potter, 1988) and that total nonresponse for the IPC next-of-kin survey (an important source of data for the residence profiles) is also associated with an absence of HCFA certification (Potter, 1989). An association with residence data completion status was not found for difficulty with bathing ($p = .051$), facility size ($p = .08$), and Medicare insurance coverage ($p = .07$).

Persons sampled as new admissions and missing only pre-admission residence data ($Partial_2$s) were similar to the $Partial_1$ group, and their respective complete counterparts, with respect to the findings for SMSA and certification status. In addition, the findings for Medicare and bathing difficulty were significant at the .05 level. However, unlike the $Partial_1$'s, the $Partial_2$'s were less likely than their complete counterparts to have private insurance coverage (38 verses 64 percent) and less likely to have difficultly dressing, bathing, walking, and transferring in and out of beds or chairs. Male gender was also found to be associated with the absence of complete residence data. The finding for private insurance coverage may be driven in part by the item nonresponse for this variable; however, the difference between the $Partial_2$'s and their complete counterparts is so large that even if all of the item nonresponse was attributed to a single category of completion status a large difference would still remain evident.

For persons sampled as admissions and in the residual category ($Partial_3$s, those sampled as an admission and missing post-discharge, or both pre-admission and post-discharge data), differences were found for private insurance coverage, and difficulty dressing and feeding in comparison to their complete counterparts.

What was surprising was the finding that family relationships -- marital status and living kin -- were not significantly different for the Partials in comparison to their complete counterparts. It was hypothesized that persons having no living kin would be more likely to have an incomplete residence profile than their complete counterparts. Especially since respondents for the IPC Next of Kin Survey were responsible for providing 76 percent of the Completes' pre-admission residence data (Potter and Cunningham, 1990). It may be that the data here are biased in this respect because persons with no kin could be more likely to be total nonrespondents or conditional nonrespondents (Table 1) and therefore excluded from these analysis.

It was also hypothesized that the Partials would significantly differ from their complete counterparts with respect to their available residence data. For example, that the Partial's would be more frequently discharged from a NH, than a complete. To investigate, three measures of 1987 nursing home utilization and a measure of residence profile were computed and expressed as means: number of unique stays in a NH, whether the sampled person ever experienced a discharge from a NH during 1987, number of days in a NH during 1987, and a measure for total number of stays in any kind of place during '87. For the completes, these measures were constructed to reflect their complete 1987 residence profile. For the Partial's, the measures were constructed to reflect only available residence data, that is their residence data prior to any imputation. Statistically significant differences between the two groups were assessed with a large sample two-sided Z test statistic. Standard errors were calculated using the program SESUDAAN (Shah, 1981), which considers the complex nature of the IPC design.

As Table 4 shows, even prior to imputation, persons with incomplete residence data had more stays in NH's, had more stays in any kind of place during the year, and were more likely to be discharged from a NH. The data illustrate that persons with partial residence data are more likely to move around during the year, both into and out of NH's providing a possible explanation for their partial residence profile status. As expected the Partials spent less time in a NH during '87 than their comparable comparison group, a finding consistent with the fact that Partial's are missing residence data.

Evaluation of Imputation Methods

Method of Imputation

A minimum distance function technique was used to impute missing residence data (Federal Committee on Statistical Methodology, 1980) to the 541 persons with missing residence profiles for 1987. This was chosen over alternative techniques because of our desire to maximize the use of available residence data for those persons requiring imputation (the Partials). This is especially important considering that the Partials are known to have significantly different patterns of NH use in comparison to their counterparts (Table 5). As part of our approach we also wanted to utilize person characteristics data known from the literature to be associated with discharge from a NH, admission to a NH, or associated with residence data completion status (as observed from these data). Our desire to utilize several variables simultaneously (particularly the residence data variables) without collapsing across cells drove us to use a minimum distance function technique. This was operationalized

using weighted sequential hot-deck imputation software (Iannachione, 1982). The weighted sequential hot-deck software was designed to impute data from individuals with complete information to individuals with missing data but with similar characteristics. Variables with known values that have been determined to be significant predictors of the measures to be imputed are used to form groups of recipients who are missing information. Within such groups, data are assigned from donors to recipients, taking into account the weights associated with each person in the complex survey. In this instance, the software was used to impute residence data from persons with complete residence data (the donors) to individuals with partial data (the recipients) while maintaining the integrity of the available partial residence data. This was done by using all available residence data as classification variables to form groups of like donors. Also used as classification variables were activities of daily living deficits, insurance status, sex, marital status, age and sample type. Because of our desire to utilize all available residence data for the partials, the donor groups formed were very small. The specifications set in the software for donor to recipient ratios and numbers of donors per group were set to a minimum. In essence, the donor became the person whose residence profile most closely resembled that of the partials' available data, while simultaneously controlling for demographic characteristics. All imputation groups created by the cross of the numerous classification variables were reviewed by hand to insure that the integrity of the available residence data was maintained. At times there was some collapsing of cells by hand across demographic items. In all, 24 imputation subgroups were utilized to maximize the use of available residence data for the partial groups (see Potter and Cunningham, 1990). Upon completion of the imputation process, 1987 residence profiles were considered complete for 5,585 persons (93 percent of the sampled population).

Results of Analytical Comparisons for NH Utilization Measures

To evaluate the results of the residence data imputation, seven NH utilization measures and two residence profile measures, expressed as means, were constructed for calendar year 1987. In addition to those measures previously described, these included two measures for admission(s) to a NH, three measures for discharge(s) from a NH, and a count of the total number of unique places of residence in any kind of place during 1987. Source data for the Partials included imputed residence data. Means for each of the measures were contrasted (using a two-sided Z test statistic) between those with no imputed data (the Completes) and to those with imputed data (the Partials). The complete group was also contrasted to the total population (the Completes and the Partials combined) to assess the effect the imputation had on mean NH utilization and residence profile estimates. Because the assumption of independence is violated for this latter comparison, the groups of complete responders were considered the population parameter for purposes of calculating the test statistic.

As Table 5 shows, the $Partial_1$s (those sampled as current residents and requiring only post-discharge imputation) were, significantly different from their counterpart completes for all but one of the nine measures even after imputation of their residence data. On the average for 1987, the $Partial_1$'s experienced fewer days in NH's, more admissions to a NH, more discharges from a NH, had more stays and stayed in more places than their complete counterparts. The $Partial_2$s and $Partial_3$s (those sampled as admissions) also spent, on the average, fewer days in a NH, and experienced more stays and stayed in more places than their counterpart completes. However, when the Partials and Completes are combined and contrasted to the complete group, virtually all significant differences for the seven NH utilization measures disappear. Significant differences do remain for the two measures of residence profile.

To determine whether the differences observed between those with complete residence data and those with partial data were the result of the residence profile imputation, utilization was further assessed for a subset of the population known to have experienced a live discharge from the sampled NH. Because of the small numbers of persons with incomplete residence data in this population, the two partial groups were combined into a single group for analysis. For persons missing some residence profile data, live discharge status was determined using only pre-imputation residence data (i.e., no imputed data were used to construct the measures). Thus, the results shown in Table 6 control for differences in residence data that occurred prior to imputation and show that after imputation persons with imputed data are no different from persons with complete residence data with respect to the measures of NH utilization and residence profile considered here. Similar comparisons were done for five additional sub-populations with similar results.

The impact of the residence profile imputation strategy on the precision of NH utilization and expenditure estimates was also assessed for both sample types as well as all NH users (the two sample types combined). Considered here are five of the NH utilization and residence profile measures considered earlier in addition to mean annual expenditures for NH use, mean per diem costs for NH use, and five measures for the percent of total NH dollars paid by: Medicaid, Medicare, family, the Veterans Administration (VA) and Supplemental Security Income (SSI). For each of the measures, the ratio of the standard errors derived from persons with complete residence data and persons with complete or partial data was computed. Controlling for the measures under consideration, the mean SE ratio for 70 some domain specific estimates (the selected characteristics presented in Table 3) were determined and are presented in Table 7. For the measures of NH use and residence profile considered, the results observed are specific to the population of interest. Among persons sampled as admissions, the inclusion of the Partials and their imputed residence data improves the precision of the standard error, on the average, by 11 percent when estimating mean number of NH utilization days. However, when estimating for the population of current residents or all NH users, the inclusion of the partials has a deleterious effect on the precision of the survey estimates (on the average, a 2 to 9 percent loss). However, regardless of the population of interest, when estimating NH expenditures and sources of payment the inclusion of the partials and their imputed residence data consistently shows a significant improvement to the precision of the survey estimates (on the average a 1 to 13 percent gain).

Summary

Data from the 1987 National Medical Expenditure Survey, Institutional Population Component were used to characterize partial and complete respondents with respect to the outcome measures based on event history data. In this analysis, data for a sample of persons who used a nursing or personal care home anytime during 1987 were used to examine the 1987 residence profiles of persons with complete residence data and contrast them to persons with some missing residence data. A brief description of the methodology used to construct the residence profiles as well as to impute missing residence data were provided. Results of the analysis undertaken to evaluate the effects of the imputation on NH utilization estimates were presented. The latter included an assessment on the precision of the survey estimates.

The results indicated that after controlling for sample type, persons who spent anytime in a nursing or personal care home during 1987 and for whom complete residence data were collected were significantly different from persons for whom only partial residence data were collected. The partial and the complete respondents differed, most notably, on their insurance coverage and health status. More importantly though, persons with missing residence data had significantly different profiles of residence history than persons with complete data, and these differences were not the result of missing data. We found that after imputation, the complete and the partial respondents continued to exhibit different patterns of residence profile and that these differences were not the result of the imputation but rather the result of their pre-imputation residence profiles (i.e., persons with missing residence data moved from place to place more often than persons with complete data). The results furthered showed that the imputation of the missing residence data, for the most part, failed to change the estimates of NH use and residence profile considered here, the most important being the estimate for total number of days of NH use.

With respect to the analysis on precision, we found that the imputation improved the precision of the estimates for total number of NH use days, by 11 percent, among those sampled as admissions but that among those sampled as current residents and for the total NH user population, the inclusion of the partial respondents and their imputed residence data had a deleterious effect on the precision of NH utilization and residence profile estimates. However, for six of the seven NH expenditure estimates considered here (including the measures for sources of payment) a modest gain to the precision of the survey estimates was observed by the inclusion of the partials and the imputed event history data.

Still to be assessed is the impact of the imputation on the magnitude of the bias to the NH expenditure estimates, if any, were we to exclude the partial respondents. We hypothesize at this time that for purposes of estimating to all NH users, the exclusion of the those with some missing residence data, after adjusting for their exclusion with a nonresponse adjustment, would make no significant differences to the estimate of total expenditures for NH use in 1987. However, when the population of interest is a subset of all NH users, such as the population of all discharges, if a bias exists, it is more likely that the estimate of NH expenditures would be slightly underestimated because persons sampled as admissions with complete residence data are less likely to be discharged than persons sampled as admissions with partial residence data. Investigation of this bias is an area of future research.

In conclusion, these data provide some evidence that by using a minimum distance function technique it is possible to impute residence data for multiple events to persons missing some event history data and in some circumstances the imputation can reduce the nonresponse bias as well as improve the precision of the survey estimates.

References

Cohen, S.B., and D.E.B. Potter. (1990a). An analysis of part-year response in the Institutional Population Component of the National Medical Expenditure Survey. Presented at the Winter Meeting of the American Statistical Association, Orlando, FL. Proceedings of the American Statistical Association, Section on Survey Research Methods. American Statistical Association, Alexandria VA.

Cohen, S.B., and D.E.B. Potter. (1990b). An estimation strategy to represent the Institutional Population Component in the 1987 National Medical Expenditure Survey. Presented at the Annual Meeting of the American Public Health Association, New York, New York.

Cohen, S.B., P. Flyer and D.E.B. Potter. (1987). Sample design of the Institutional Population Component of the National Medical Expenditure Survey. Presented at the Annual Meeting of the American Public Health Association, New Orleans, Louisiana.

Cox, D.R. (1975). Partial likelihood. Biometrika 62: 269-76.

Edwards, W.S., and B. Edwards. (1989). National Medical Expenditure Survey: Questionnaires and Data Collection Methods for the Institutional Population Component, Methods 1. Department of Health and Human Services, Public Health Service, National Center for Health Services Research and Health Care Technology Assessment, DHHS Pub. No. (PHS) 89-3440.

Federal Committee on Statistical Methodology, Subcommittee on Matching Techniques. (1980). Statistical Policy Working Paper 5: Report on Exact and Statistical Matching Techniques. U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, Washington, D.C., pp. 41-45.

Iannachione, V.G. (1982). Weighted sequential hot-deck imputation macros. Proceedings of the Seventh Annual SAS User's Group International Conference. SAS Institute, Inc. Cary, N.C. pp. 759-763.

Petersen, T. (1991). Guest editor's introduction to the special issue on event history analysis. Sociological Methods and Research, 19:3 p. 267.

Potter, D.E.B. (1988). Nonresponse in a survey of nursing homes and facilities for the mentally retarded: Findings

from the National Medical Expenditure Survey, Institutional Population Component. Proceedings of the American Statistical Association, Section on Survey Research Methods. American Statistical Association, Alexandria VA.

Potter, D.E.B. (1989). Nonresponse in a survey of nursing home residents. Proceedings of the American Statistical Association, Section on Survey Research Methods. American Statistical Association, Alexandria VA.

Potter, D.E.B. and P.J. Cunningham. (1990). Constructing year-long profiles of residence histories for a national sample of nursing home users. Proceedings of the American Statistical Association, Section on Survey Research Methods. American Statistical Association, Alexandria VA.

Shah, B.V. (1981). SESUDAAN: Standard Errors Program for Computing Standardized Rates from Sample Data.Technical Report. Research Triangle Institute, Research Triangle Park, North Carolina.

Shah, B.V., L.M.LaVange, B.G. Barnwell, J. E. Killinger, and S.E. Wheeless. (1989). SUDAAN Professional Software for SUrvey DAta ANalysis: Procedures for Descriptive Statistics User's Guide. Research Triangle Institute, Research Triangle Park, North Carolina.

Readers who wish to obtain a complete set of tables and figures presented in this paper should contact D.E.B. Potter, AHCPR, at the above address.

The views expressed in this paper are those of the authors. No official endorsement by either the Agency for Health Care Policy and Research or the Department of Health and Human services is intended or should be inferred.

The authors acknowledge the invaluable computer processing support of Kathy McMillan, Ase Sewall and Maria Nicholson of Social and Scientific Systems, Inc.