

## Multiple Causes of Death for the National Health Interview Survey

*John Horm, National Center for Health Statistics*

---

### **Abstract**

*The National Health Interview Survey (NHIS) is a nationally representative health survey of the United States population. The NHIS is a rich resource for national and subnational health information such as chronic and acute conditions, doctor visits, hospital stays and a wide variety of special health topics knowledge, attitudes, and behaviors each year. Basic socio-demographic information is routinely collected on each person in the NHIS. The NDI contains records for virtually 100 percent of persons who die in the United States. Respondents to the NHIS who are age 18 or over are now routinely linked with the National Death Index (NDI) to create a new resource of immense public health and epidemiologic potential. An automated probabilistic approach has been used to link the two data files from the date of interview through 1995 and classify the linked records as either true (deceased) or false (alive) matches. It is estimated that over 97 percent of deceased persons and 99 percent of living persons are correctly classified as to vital status. The linked NHIS-NDI files contain all of the survey information along with vital status, multiple causes of death and date of death if deceased.*

### Introduction

The National Health Interview Survey (NHIS) is a large in-person health survey of the United States population conducted annually by the National Center for Health Statistics (Dawson and Adams, 1987).

Health and health-related information is collected on approximately 122,000 persons per year (42,000 households) among the civilian, non-institutionalized population (note that since matching with the NDI is done only for persons aged 18 and over, the sample size for this purpose is about 85,000 persons). The NHIS consists of a basic health and demographic questionnaire (BHD) with information on every person in the household. The BHD contains basic socio-demographic information, acute and chronic conditions, doctor visits, hospital stays, and related items. In addition to the BHD, one or more surveys on Current Health Topics (CHT) is also conducted each year. The CHT surveys are usually administered to one randomly selected sample person over the age of 18 in each family although there are some family-style CHT surveys. The sample-person CHT surveys yield information on about 42,000 persons per year. Recent CHT surveys include the following content areas: alcohol use; cancer epidemiology and control; child health; health insurance; adult immunization; Year 1990 health objectives; Year 2000 health objectives and others. All questionnaires and topic areas included from 1985 through 1989 have been published by Chyba and Washington (1993). Response rates for both components of the NHIS are high: 95 percent for the BHD and about 85 percent for the CHT's.

The NDI is a central computerized index with a standard set of identifying information on virtually every decedent in the United States since 1979 (Boyle and Decoufle, 1990) managed by the National Center for Health Statistics and can be used to enumerate and identify decedents in epidemiologic studies. The NDI produces matches between user records and death records based on a set of twelve criteria. The user must then develop a methodology to classify the potential matches returned by the NDI as either true or false matches.

The approach taken here to classify the NHIS-NDI potential matches is a modification of the probabilistic approaches developed by Fellegi and Sunter (1969) and refined by Rogot, Sorlie, and Johnson (1986).

## Methods

The NDI contains records on all deaths occurring in the United States since 1979 and is fully documented in the National Death Index User's Manual (1990). The NDI has developed a set of 12 criteria under which matches between user records and NDI records are produced. These criteria are based on various combinations of Social Security Number, date of birth, first name, middle initial, and last name. The 12 matching criteria are:

- Social security number and first name;
- Social security number and last name;
- Social security number and father's surname;
- If female, Social security number, last name (user's record) and father's surname (NDI record);
- Month and year of birth and first and last name;
- Month and year of birth and father's surname;
- If female, month and year of birth, first name, last name (user's record) and father's surname (NDI record);
- Month and year of birth, first and middle initials, and last name;
- Month and  $\pm 1$  year of birth, first and middle initials, and last name;
- Month and  $\pm 1$  year of birth, first and last names;
- Month and day of birth, first and last names; and
- Month and day of birth, first and middle initials, and last name.

An NDI record is matched to a user record if any one of the above 12 criteria result in a match.

An indication of agreement between the user record and the NDI record is returned to the user for each of the seven items involved in the twelve matching criteria. In addition to the items involved in the matching criteria the NDI returns an indication of agreement/disagreement between the user record and the NDI record on five additional items: age at death; race; marital status; state of residence; and state of birth. Multiple NDI records may be matched to a single user record and a possibly large number of false positive matches may be returned by the NDI. Matches between NDI records and NHIS records are referred to as *potential matches*.

The NHIS routinely collects all of the seven data items used by the NDI for matching as well as the five additional items used for assessing the quality of potential matches. The NHIS has essentially 100 percent complete reporting of these items except for social security number (SSN) and middle initial. Completeness of reporting of SSN and middle initial varies by year but is generally between 65 and 75 percent. Various studies have indicated that the NDI is capable of identifying over 90 percent of known deaths (Patterson and Bilgrad, 1986; Stampfer et al., 1984; Williams, Demitrack and Fries, 1992) with some studies finding that the proportion is in the upper 90's when a full set of identifiers is available (Calle and Terrell, 1993; Curb et al., 1985; Horm and Wright, 1993). Social Security Number is a key identifier in the matching process. When the SSN is not available the proportion of known deaths identified drops to about 90 percent.

Tepping (1968) developed a model for computerized matching of records from the perspective of the cost of making correct or incorrect decisions about potential matches. Fellegi and Sunter (1969) developed a theory-based approach for record linkage which incorporated the concept of weighting factors with the weight being positive if the factor agreed and negative if it disagreed. With the magnitude of the weight being inversely proportional to the frequency of the factor in the population. This approach was refined by Rogot, Sorlie, and Johnson (1986) who used binit weights [ $\text{Log}_2(1/p_i)$ ] where  $p_i$  is the proportion of the population with the  $i^{\text{th}}$  characteristic. Newcombe, Fair, and Lalonde (1992) while not espousing a particular form for the weights did make a case for the necessity of weighting by something more than simple agreement/disagreement weights.

## Weights

Weights for each of the eleven items used for assessing the quality of the potential matches were constructed based on the composition of the 1988-91 NHIS and 1986-91 U. S. deaths (SSN is handled separately).

A weight is the base 2 logarithm of the inverse of the probability of occurrence of the characteristic based on the above files. For example, since males constitute about 46.3 percent of the population aged 18 and over, the weight is  $\log_2(1/.463) = 1.11$ . Weights are constructed in a similar manner for race, last name, father's surname, birth month, day, and year, state of residence, and state of birth. Since middle initials are sex-specific, sex-specific weights were constructed for middle initial. Weights for marital status were constructed to be jointly age and sex specific. First name weights are both sex and birth year cohort (<1926, 1926-1935, 1936-1955, and >1955) specific because of secular trends in the assignment of first names.

Weights may be either positive or negative. If a particular item matches between the NHIS record and the NDI record, the weight is positive. If the item does not match, the weight is negative. Weights for items missing from the NHIS file, the NDI file, or both are assigned a weight of zero.

Last name weights have been modified for females. Since some females change their surnames upon marriage, divorce, remarriage, etc., matching on surname only may produce false non-matches. The NDI returns an indication of a match on the father's surname as well as last name which is used as auxiliary information for females. If last name does not match on the two records (the last name weight is negative), the last name weight is replaced with the father's surname weight if positive, otherwise the last name weight is retained. This approach provided the best classification performance for females.

Because all information provided to the NDI is proxy reported and information provided to the NHIS may be proxy reported, there is a considerably likelihood that one of the two files may contain a respondent's given first name while the other contains his/her commonly used nickname. We have constructed files of common nicknames which are used in the classification process if the first name on file does not provide a good match.

Frequency-based weighting schemes such as proposed by Fellegi and Sunter and Rogot, Sorlie, and Johnson are attractive since the rarer occurrences of a matching item is given more weight than more common occurrences. However, the user is still left with the problem of properly classifying matched records into at least minimal categories of true matches, false matches, and questionable matches. Recent work by Belin (1993) and Belin and Rubin (1993) suggests that the false-match rate is sensitive to the setting of cut-points.

## **Calibration Samples**

Calibration samples need to have known vital status information such as date and location of death, and ideally, death certificate number on the sample subjects based on sources independent of the NDI. Two NCHS surveys meet this criteria.

The 14,407 persons who participated in the NHANES I examination survey (1971-75) were used as the first calibration sample. Active followup was conducted on this sample to ascertain the vital status of the participants and death certificates obtained for persons found to be deceased (Finucane et al., 1990). NHANES is a large nationally representative survey and is sufficiently similar to the NHIS to be used as a calibration sample for developing a methodology for classification of the NHIS-NDI matches.

The NHANES I followup sample was then matched to the NDI and randomly stratified into two samples, a developmental sample and a confirmation sample.

Any one calibration sample may have an inherent structural process which differs systematically from the

target sample. Even though the NHANES sample was randomly stratified into two samples, systematic differences between NHANES and the NHIS could exist in both parts. Thus a second calibration sample was used to counteract potential structural differences. The second calibration sample used was the Longitudinal Study on Aging (LSOA) (Kovar, Fitti, and Chyba, 1992), a subset of the 1984 NHIS. The data used from this sample were those participants aged 70 and over at the time of interview and followed through August, 1988. Vital status was obtained independent of the NDI by interviewer followback in both 1986 and 1988.

### Classification of Potential Matches

Potential matches returned by the NDI must be classified into either true or false matches. This is done by assigning a score, the sum of the weights, to each match.

$$\begin{aligned} \text{Score} = & W_{\text{firstname}} \times \text{sex} \times \text{birthcohort} + W_{\text{middleinitial}} \times \text{sex} + W_{\text{lastname}} \\ & + W_{\text{race}} + W_{\text{maritalstatus}} \times \text{sex} \times \text{age} + W_{\text{birthday}} \\ & + W_{\text{birthmonth}} + W_{\text{birthyear}} + W_{\text{stateofbirth}} + W_{\text{stateofresidence}} \cdot \end{aligned}$$

The NHANES I developmental sample suggested that classification efficiency could be increased by grouping the potential matches into one of five mutually exclusive classes based on which items matched and the number of items matching. These classes are:

- Class 1: Exact match on SSN, first, middle, and last names, sex, state of birth, birth month and birth year.
- Class 2: Exact match on SSN but some of the other items from Class 1 do not match although certain cases were moved from Class 2 to Class 5 because of indications that the reported SSN belonged to the spouse.
- Class 3: SSN unknown but eight or more of first name, middle initial, last name, birth day, birth month, birth year, sex, race, marital status, or state of birth match.
- Class 4: Same as Class 3 but less than eight items match.
- Class 5: SSN known but doesn't match. Some cases were moved from Class 5 to Class 3 because of indications that the reported SSN belonged to the spouse.

In this classification scheme all of Class 1 are considered to be true matches implying that the individuals are deceased while all of the Class 5 matches are considered false matches. Assignment of records falling into one of Classes 2, 3, or 4, as either true matches or false matches was made based on the score and cut-off points within class. Records with scores greater than the cut-off scores are considered true matches while records with scores lower than the cut-off scores are considered false matches.

The cut-off scores were determined from the NHANES I developmental sample using a logistic model. The logistic model was used within each of classes 2, 3, and 4 to determine cut-off scores in such a manner as to jointly maximize the *number and proportion* of records correctly classified while minimizing the *number and proportion* of records incorrectly classified. The cut-off scores were then applied to the NHANES I confirmation sample for refinement. Slight fine-tuning of the cut-off scores was required at this stage because of the relatively small sample sizes. Finally the weights and cut-off scores were applied to the LSOA sample for final confirmation. Further refinements to the cut-off scores were not made.

### Results

The recommended cut-off scores are estimated to correctly classify over 97 percent of NHIS decedents and over 99 percent of living persons. It is known that the NDI misses about five percent of known decedents. An adjustment for this has not been included in these classification rates.

### Subgroup Biases in Classification

The correct classification rate for females who were known to be deceased is about 2.5 percentage points poorer for females than males. This is due to linkage problems caused by changing surnames through marriages, divorces, and widowhood. Even though father's surname is being used to provide additional information there still remain problems of correctly reporting and recording surnames in both the survey and on the death certificates. Both males and females have the same correct classification rates for living persons.

Among non-whites there are multiple problems including lower reporting of social security numbers and incorrect spelling/recording of ethnic names. The correct classification rates for non-white decedents dropped to 86 percent while the classification rate for living persons remained high at over 99 percent. The classification rate for deceased non-white females was about three percent lower than that for non-white male decedents (84.7 percent and 87.8 percent, respectively). These biases are due to the relatively large proportions of non-white decedents in Class 4 because of incorrect matching information. Females and non-whites falling into Classes 1, 2, 3, or 5 have the same classification rates as white males.

### Discussion

Application of the above outlined matching and classification methodology to 1986 through 1994 NHIS survey year respondents provides death follow-up from the date of interview through 1995. The linkage of these files yields approximately 900 deaths for each survey year for each year of follow-up. For example, there are 7,555 deaths among respondents to the 1987 survey with an average of 8½ years of follow-up. Although years can be combined to increase the sample sizes for data items included in the NHIS core (BHD items), this is not generally the case for supplements which change topic areas each year. NHIS supplements are usually administered to one randomly chosen person age 18 or over in each household. This results in an annual sample size for the NHIS of about 42,000 persons. The number of deaths among such supplement respondents would be approximately one-half the number of deaths listed above (e.g., about 450 deaths per survey year per year of follow-up).

The NHIS-NDI linked files (NHIS Multiple Cause of Death Files) can be used to estimate mortality rates (although caution must be given to biases), life expectancies, and relative risks or odds ratios of death for a wide variety of risk factors while controlling for the influence of covariates. For example, the impact of poverty or health insurance status on the risk of dying could be explored while simultaneously controlling for age, sex, race, acute or chronic conditions. Or, mortality rates according to industry or occupation could be developed or for central city residents relative to rural residents. Such analyses are possible because the NHIS carries its own denominators (number at risk).

### References

- Belin, T.R., and Rubin, D.B. (1995). A Method for Calibrating False-Match Rates in Record Linkage, *Journal of the American Statistical Association*, 90, 430, 694-707.
- Belin, T.R. (1993). Evaluation of Sources of Variation in Record Linkage Through a Factorial Experiment, *Survey Methodology*, 19, 1, 13-29.
- Boyle, C.A., and Decoufle, P. (1990). National Sources of Vital Status Information: Extent of Coverage and Possible Selectivity in Reporting, *American Journal of Epidemiology*, 131, 160-168.
-

- Calle, E. E., and Terrell, D.D. (1993). Utility of the National Death Index for Ascertainment of Mortality among Cancer Prevention Study II Participants, *American Journal of Epidemiology*, Vol 137, 235-241.
- Chyba, M.M., and Washington, L.R. (1993). Questionnaires from the National Health Interview Survey, 1985-89, National Center for Health Statistics, *Vital and Health Statistics*, 1(31), DHHS Publication No. (PHS) 93-1307, Public Health Service, Washington, D.C. U.S. Government Printing Office.
- Curb, J.D.; Ford, C.E.; Pressel, S.; Palmer, M.; Babcock, C.; and Hawkins, C.M. (1985). Ascertainment of Vital Status Through the National Death Index and the Social Security Administration, *American Journal of Epidemiology*, 121, 754-766.
- Dawson, D.A., and Adams, P.F. (1987). Current Estimates from the National Health Interview Survey, United States, 1986, National Center for Health Statistics, *Vital and Health Statistics*, Series 10, No. 164, DHHS Pub. No. (PHS) 87-1592, Public Health Service, Washington, D.C. U.S. Government Printing Office.
- Fellegi, I.P., and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.
- Finucane, F.F.; Freid, V.M.; Madans, J.H.; Cox, M.A.; Kleinman, J.C.; Rothwell, S.T.; Barbano, H.E.; and Feldman, J.J. (1990). Plan and Operation of the NHANES I Epidemiologic Followup Study, 1986, National Center for Health Statistics, *Vital and Health Statistics*, Series 1, No. 25, DHHS Pub. No. (PHS) 90-1307, Public Health Service, Washington, D. C. U.S. Government Printing Office.
- Horm, J.W., and Wright, R.A. (1993). A New National Source of Health and Mortality Information in the United States, *Proceedings of the Social Statistics Section, American Statistical Association*, San Francisco.
- Kovar, M.G.; Fitti, J.E.; and Chyba, M.M. (1992). The Longitudinal Study on Aging: 1984-90. National Center for Health Statistics, *Vital and Health Statistics*, 1(28).
- National Center for Health Statistics (1990). *National Death Index User's Manual*, U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics, DHHS Pub. No. (PHS) 90-1148.
- Newcombe, H.B.; Fair, M.E.; and Lalonde, P. (1992). The Use of Names for Linking Personal Records. *Journal of the American Statistical Association*, 87, 1193-1208.
- Patterson, B.H., and Bilgrad, R. (1986). Use of the National Death Index in Cancer Studies, *Journal of the National Cancer Institute*, 77, 877-881.
- Rogot, E.; Sorlie, P.; and Johnson, N.J. (1986). Probabilistic Methods in Matching Census Samples to the National Death Index, *Journal of Chronic Diseases*, 39, 719-734.
- Stampfer, M.J.; Willett, W.C.; Speizer, F.E.; Dysert, D.C.; Lipnick, R.; Rosner, B.; and Hennekens, C.H. (1984). Test of the National Death Index, *American Journal of Epidemiology*, 119, 837-839.
- Tepping, B.J., (1968). A Model for Optimum Linkage of Records, *Journal of the American Statistical Association*, 63, 1321-1332.
- Williams, B.C.; Demitrack, L.B.; and Fries, B.E. (1992). The Accuracy of the National Death Index When Personal Identifiers Other than Social Security Number Are Used, *American Journal of Public Health*,

82, 1145-1147.