Suggested citation: National Center for Health Statistics. Office of Analysis and Epidemiology, 2004 National Nursing Home Survey (NNHS) Linked Mortality File, mortality follow-up through 2006: Matching Methodology September 2009. Hyattsville, Maryland. (Available at the following address: http://www.cdc.gov/nchs/data/datalinkage/matching_methodology_04nnhs_final.pdf)

The 2004 National Nursing Home Survey (NNHS) Linked Mortality File: Mortality follow-up through 2006

Matching Methodology

Introduction

The 2004 National Nursing Home Survey (NNHS) Linked Mortality file includes mortality status ascertained primarily through probabilistic record matching to the National Death Index (NDI) through December 31, 2006. The NDI record linkage employs a matching methodology that is similar, but not identical, to the standard methodology offered by the NDI. However other sources of mortality information may be utilized to determine vital status[1].The 2004 NNHS Linked Mortality file provides researchers the opportunity to conduct studies examining profiles of long term care utilization and subsequent mortality.

The National Death Index Matching Algorithm

The NDI is a NCHS centralized database of all U.S. deaths beginning in 1979 that is used to match a NCHS survey record to a NDI record of death according to NDI matching criteria. The NDI contains the following identifying information for each death, which are used for matching purposes:

Social Security number
First name
Middle initial
Last name
Month of birth
Day of birth
Year of birth
Sex
Father's surname
State of birth
Race
State of residence
Marital status

NCHS prepares records[2] of the survey participant based upon these identifiers that are submitted to the NDI and used in the matching process. The 2004 NNHS study collected most of the data items used by the NDI for matching (Table 1). Before the NDI processes any submission record, each record is screened to determine if it contains at least one of the following combinations of identifying data elements:

1. Social Security number, sex, full date of birth present
2. Last name, first initial, month of birth, year of birth present

[1] Please refer to the information on mortality source in the Analytic Guidelines.
[2] The NDI allows multiple submission records for each survey participant and NDI records can be matched to any or all of the submission records created for a survey participant. Appendix A describes scenarios under which NCHS generates alternate submission records.

3. Last name, first initial, Social Security number present

Any survey participant submission record that did not meet these minimum data requirements was ineligible for record linkage. Table 2 lists the number of survey participants and eligibility status. The NDI system selects death record matches based on a set of established match criteria. The seven criteria listed below were the criteria in use at the time of the 2004 NNHS-NDI match.

1. Social Security number
2. First and last name, exact month of birth, year of birth within 1 year
3. Last name, first initial and middle initial, exact month of birth, year of birth within 1 year
4. First and last name, exact month of birth, exact day of birth
5. Last name, first initial and middle initial, exact month of birth, exact day of birth
6. First name, father's surname, exact month of birth, exact year of birth
7. For females only, first name, exact month and year of birth, and last name from the user's record matching birth surname on the NDI record (for females who change their name after marriage, but don't supply a birth surname)

Agreement on names may be based upon exact spelling matches or, since spelling variants of names are common, based upon the way a name sounds rather than how it is spelled[3]. Any NDI record that matches a 2004 NNHS submission record on any one of these seven criteria is selected as a potential match. As one or more NDI records may be matched to a given 2004 NNHS submission record, the NDI record selection process can return several potential matches for each 2004 NNHS person, many of which will be non-matches or duplicate records. Users interested in a detailed description of the standard NDI matching methodology should refer to the NDI.

Scoring and classifying potential match records

As previously described, there are seven ways that a 2004 NNHS survey participant submission record may match to a NDI record. For each potential match, the NDI provides a code indicating whether there is agreement, disagreement, or no basis for comparison for each identifier. NCHS assigns a score to each potential match reflecting the degree of agreement between the identifying information on the 2004 NNHS submission record and the NDI death record. The score is based upon probabilistic weights assigned to each of the identifying data items used in the 2004 NNHS-NDI record match[4]. For example, a common first name, such as "John", that has a higher probability of occurrence in the population has a lower weight than an uncommon name such as "Jonas". Weights are either positive or negative. If there is agreement between the 2004 NNHS record and the NDI record for a particular identifying data item, the weight is positive. If there is no agreement, the weight is negative. With the exception of middle initial, data items that

[3] The sound alike system is a variation of the New York State Identification Intelligence System or NYSIIS, which converts a name to a phonetic coding. For example, records with last names Smith and Smyth receive equivalent NYSIIS codes and both would be selected as a potential match for a 2004 NNHS submission with Smith (or Smyth) as a last name.

[4] NCHS developed a new system of assigning weights to each identifying data element that is similar, but not identical, to the standard methodology offered by the NDI. NCHS developed the weights, known as binit weights, based upon the frequency of occurrence of the 12 data items in the NDI files for years 1979 to 2000, which represents about 49 million persons. The weights correspond to $[\mathrm{Log}_2 (1/p_i)]$: the base 2 logarithm of the inverse of the probability of occurrence of the value of the identifying data item on the submission record.

are missing on the 2004 NNHS submission record, the NDI record, or both receive a weight of zero. The score for each potential match is the sum of the weights for each individual data item.

$$\textbf{\textit{Score}} = \{\sum W_{SSN1} + \ldots + W_{SSN9}{}^{5}\} + W_{firstname\ x\ sex\ x\ birthyear} + W_{middleinitial\ x\ sex} + W_{lastname} + W_{race} + W_{sex} + W_{maritalstatus\ x\ sex\ x\ age} + W_{birthdate} + W_{birthmonth} + W_{birthyear} + W_{stateofbirth} + W_{stateof\ residence}$$

After scoring the potential matches, each is categorized into one of five mutually exclusive classes. Whereas weighting and scoring take into account the probability that the 2004 NNHS record and the NDI record share a particular value for the identifying items, the classes take into account which identifying items agree. They reflect the fact that some of the 12 NDI identifying items are more important for determining true matches than others (e.g. SSN versus state of birth) and that non-changing identifying information is more important than information that can change over time (e.g. birth surname versus marital status).

As SSN is a key identifier in the matching process, each 2004 NNHS-NDI record match is initially classified according to whether SSN is present and agrees (Class 1 or 2), is present but disagrees (Class 5) or is missing (Class 3 or 4). The final five classes used by NCHS for the 2004 NNHS Linked Mortality file are as follows.

Class 1: Agrees on at least 8 (of 9) digits of SSN, first name (including NYSIIS match), middle initial (including blank), last name (including NYSIIS match), birth year (+/- 3 years), birth month, sex, and state of birth.

Class 2: Agrees on at least 7 (of 9) digits of SSN and at least 5 more of the following items: first name (including NYSIIS match), middle initial (including blank), last name (including NYSIIS match), birth year (+/- 3 years), birth month, sex, and state of birth.

Class 3: There are two types of Class 3 matches:

Type A: SSN is unknown, but last name matches (including NYSIIS match) and at least 7 of the following items agree: first name (including NYSIIS match), middle initial (including blank), last name (including NYSIIS match), birth year (+/- 3 years), birth day, sex, race, marital status and state of birth.

Type B: Records in this category were initially put in Class 5 but switched to Class 3[6]. SSN is known but 3 or more digits do not agree, but at least 8 of the following items agree: first name, middle initial (including blank), last name, birth year, birth day, sex, race, marital status, and state of birth.

[5] For a record to be assigned the maximum weight for SSN, there needs to be agreement on at least 8 digits. If seven digits agree, then 7/9 of the total weight is assigned. If fewer than seven digits agree then the total SSN weight becomes negative.

[6] This class switch occurs if after review, there is the possibility that SSN was either recorded incorrectly or that the spouse's SSN was recorded instead of the subject's SSN. All total scores were adjusted to reflect the final class code for the potential matches. For example, any record that was switched from Class 5 to Class 3 had its score adjusted to reflect that SSN is missing, with the value of 0 assigned to SSN.

Class 4: SSN is unknown on either the 2004 NNHS submission record or the NDI record and fewer than 8 of the items listed in Class 3 match.

Class 5: SSN is present but fewer than 7 (of 9) digits on SSN agree.

Selecting matches and assigning vital status

Since each eligible 2004 NNHS participant may have multiple submission records and each submission record may return one or more potential matches to a NDI record, NCHS employed a strategy to provide the single best NDI match record for inclusion on the linked mortality file.

First, 2004 NNHS-NDI potential match records that had a date of death prior to the date of interview, a score of zero or less, or final categorization of Class 5 were considered false matches and eliminated from the pool of potential matches. Next, among the remaining pool of potential matches, duplicates (i.e. match records that referred to the same death certificate) were eliminated. Many participants, however, still had more than one NDI record as a potential match. The remaining potential matches were ranked first on class (from 1 to 4) and then within class by highest score. The NDI match with the highest score within the best class was selected as the single best record match. In the event of a tie among NDI record matches for a particular 2004 NNHS record, the record underwent manual review with the tiebreaker reflecting the importance of matching items.

Next, NCHS determined whether each best record match was true or false. A true match reflects *both* the correct vital status of the survey participant and a match to the correct death certificate data. All Class 1 match records were considered true matches. For match records with Classes 2, 3, and 4, NCHS determined whether the match was true or false using cut-off scores developed from the NHANES I Epidemiologic Follow-up Survey (NHEFS) calibration sample. Within each class, matches with a score *greater than or equal* to the cut-off score were considered true matches, while records with a score less than the cut-off were considered false matches. *The cut-off scores for Classes 2, 3, and 4 were 47, 45, and 40, respectively.* In general, the process was to select the cut-off scores within Classes 2, 3, and 4 that simultaneously maximized the proportion of people correctly classified and minimized the number of people incorrectly classified, with particular attention given to minimizing the number of false positives. In addition, for a small percentage of 2004 NNHS participants, NCHS conducted a manual review of the 2004 NNHS submission record and the corresponding NDI potential matches to determine vital status.

Table 1. Percent of missing data for eligible respondents by sex: 2004 NNHS

Personal Identifying Information	**Male**	**Female**
Social Security Number [1]	0.88	0.95
First Name [1]	0.26	0.20
Middle Name [1,2]	96.70	96.72
Last Name [1]	0.23	0.17
Date of Birth [1]	0.08	0.04
Race	0.00	0.00
Sex	0.00	0.00

[1]Social Security number, Name, Date of Birth are used to select potential match records. All variables listed are utilized by the probabilistic scoring algorithm.
[2]Blank middle name/initial is considered a valid code.

Table 2. Eligibility status and NDI match status: 2004 NNHS

	Eligibility Status for Mortality Follow-up		
Sample Total	Eligible	Ineligible, insufficient data	Assumed Deceased
13,507	13,464	43	6,767

Appendix A

Creating Alternate Submission Records

The primary purpose of using alternate submission records is to increase the chances of returning a correct death record for those 2004 NNHS participants who are, in fact, deceased. The NDI allows multiple alternate submission records for each survey person.

Alternate submission records may be created for several reasons. For example, if a SSN is present but additional information indicates that the SSN is not valid, an alternate submission record will be created that does not include SSN. Name inaccuracies are the most common type of mismatch error encountered when matching to the NDI system, e.g. reporting a nickname like "Beth" for a formal name like "Elizabeth" or the presence of multi-part first or last names. In these cases, alternate submission records are created that take into account nicknames being listed as the first name, using a nickname to proper name conversion process or that use all of the components of multi-part names both separately and together.

The rules for alternate submission record creation are multiplicative in nature. For example, a participant may have both an imputed month of birth (12 separate records) and two-part first name (3 separate records) resulting in 36 NDI submission records.

Appendix B

Altering the criteria to assign vital status

The 2004 NNHS Linked Mortality file includes the NCHS recommended vital status ascertainment (MORTSTAT) for each eligible 2004 NNHS participant. NCHS also has a set of special request variables, NDI probabilistic match score (SCORE) and match criteria classification (CLASS), that users can request to alter the criteria for determining vital status and conduct their own sensitivity analyses. Below are two examples of studies using the NHIS Linked Mortality files that evaluate different criteria to ascertain vital status.

Using the 1986-1990 NHIS linked to the NDI with mortality follow-up through 1991, Liao et al. (1998) evaluated death rates using three different criteria to identify deaths. Criterion 1 was the most conservative, requiring an exact match on SSN. Only NHIS participants with a NDI record with a Class 1 or 2 match were considered true matches and assumed deceased; Criterion 2 was the NCHS recommended ascertainment of vital status using NCHS's cut-off scores[7]; and Criterion 3 was the least stringent with all of Class 1, 2, and 3 matches plus Class 4 matches with scores higher than the recommended cut-off considered true matches. Mortality estimates were lowest based upon criterion 1 and highest with criterion 3. Furthermore, the use of different criteria to determine vital status had differential effects upon death rates for sex and race/ethnic groups.

Using a prior data release of the NHIS Linked Mortality files for the NHIS years 1988-1994, NCHS conducted its own evaluation of alternative criteria to assign vital status and its impact on death rates. The analysis evaluated four criteria that differed in the established cut-off scores for determining which Class 2, 3, or 4 matches were true matches. For all four criteria, Class 1 match records were considered true matches. Criterion 1 reflected the NCHS recommended cut-off scores; criterion 2 made all Class 2 matches true matches and all Class 4 matches non-matches; criterion 3 increased by four points the cut-off scores for Classes 2, 3, and 4 from the NCHS recommended scores; and criterion 4 lowered by four points the cut-off scores for Classes 2, 3, and 4 from the NCHS recommended scores. Table 3 displays the results.

Across the four criteria, the number of Class 1 matches (assumed deceased) was 36,345. Altering the cut-off scores for the other classes, Criterion 3 produced the most conservative results with 58,934 deaths overall and criterion 4 the least conservative with 64,142 deaths. The NCHS recommended cut-off scores produced 61,021 deaths overall. Not surprisingly, altering the matching criteria differentially affected the mortality rates by race/ethnicity, sex, and age, with Hispanics being the most affected (data not shown).

Inferences from studies that use the NHIS Linked mortality files could be affected by the matching criteria chosen and the ascertainment of vital status. Researchers interested in examining the robustness of their findings can perform sensitivity analyses by altering the

[7] Users should note that the NCHS recommended classifications and cut-off scores used in this analysis do not reflect the current classifications and cut-off scores used to determine true match status between NCHS survey records and NDI records.

criteria for matching. This may be particularly important for studies examining mortality patterns for specific race/ethnic groups.

Table 3: Number of deaths assigned to NHIS participants based upon four criteria for determining NHIS-NDI record linkages to be matches or non-matches

	Match Record Class	Match Record Score	No. Assumed Dead
Criterion 1: NCHS recommendation	2 3 4	≥ 47 ≥ 45 ≥ 40	10,273 11,855 2,548
Criterion 2	2 3 4	All ≥ 45 None	10,734 11,855 --------
Criterion 3 (4 points above NCHS recommended cut-off scores)	2 3 4	≥ 51 ≥ 49 ≥ 44	10,038 10,817 1,291
Criterion 4 (4 points below NCHS recommended cut-off scores)	2 3 4	≥ 43 ≥ 41 ≥ 36	10,415 12,595 4,787