NCHS 2011 Linked Mortality Files Matching Methodology

Suggested citation: National Center for Health Statistics. Office of Analysis and Epidemiology, NCHS 2011 Linked Mortality Files Matching Methodology, September, 2013. Hyattsville, Maryland. Available at the following address:
http://www.cdc.gov/nchs/data_access/data_linkage/mortality/linkage_methods_analytical_support/2011_linked_mortality_file_matching_methodology.pdf

The National Center for Health Statistics Linked Mortality Files:
Mortality follow-up through 2011

Matching Methodology

Introduction

The NCHS Linked Mortality Files include the following surveys:

National Health Interview Survey (NHIS): 1985–2009

Continuous NHANES: 1999 – 2010

National Health and Nutrition Examination Survey (NHANES) I Epidemiologic Follow-up Study (NHEFS)

NHANES II

NHANES III (1988-1994)

2007 National Home and Hospice Care Survey (NHHCS)

1995 National Nursing Home Survey (NNHS)

1997 National Nursing Home Survey (NNHS)

2004 National Nursing Home Survey (NNHS)

The Supplement on Aging (SOA) (1984 NHIS supplement)

Second Longitudinal Study of Aging (LSOA II)

Mortality status for NCHS survey participants was ascertained primarily through probabilistic record matching with the National Death Index (NDI). The NDI is a NCHS centralized database of all U.S. deaths beginning in 1979. NCHS Special Projects Branch (SPB) employed a matching methodology for the 2011 Linked Mortality Files that was similar, but not identical, to the standard methodology currently offered by the NDI. SPB also relied on other sources of mortality information to determine vital status, such as through linkages with the Social Security Administration or through active follow-up. NHEFS, NHANES II, LSOA II, and the 1985 NNHS all contain death information ascertained during follow-up periods after the surveys. For more detail on other sources of mortality, please refer to the Data Dictionary for the 2011 Linked Mortality Files.

Changes from Prior Releases

The previous NCHS Linked Mortality Files provided mortality data through December 31, 2006. The updated 2011 Linked Mortality Files supersede previous linkages of NCHS surveys to the NDI.

This 2011 version of the NCHS Linked Mortality Files is different from the previous version in several ways: (1) additional surveys and survey years have been added; (2) eligible participants

have death information added for five additional years (through December 31, 2011) of mortality follow-up; and (3) several minor modifications were made to the matching algorithm which may have led to differences in eligibility status, vital status, or death information from the previous version; (4) beginning in 2007, the NHIS began collecting just the last 4 digits of Social Security Number, which required minor modifications to the matching algorithm that are specific to these surveys.

Linkage Fields

NCHS survey records were matched with NDI records using the following identifying information from NDI, as available, for each death:

Social Security Number (SSN)
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

- The NHIS has routinely collected all of the data items used by the NDI for matching. However, beginning in 2007 NHIS only collects the last 4 digits of the SSN and not the full nine digits.
- NHEFS, NHANES III, continuous NHANES, LSOA II, and NHHCS 2007 collected all of the data items used by the NDI for matching.
- NHANES II did not collect SSN, but SSN was obtained through secondary data collection.
- NNHS and SOA collected most of the data items used by the NDI for matching (SSN, first name, middle name, last name, date of birth, race, and sex).

NCHS SPB prepared records of the survey participants based upon the same identifiers, as available from each survey, and used them to link to death records in the NDI. To increase the likelihood of finding a match, SPB created multiple submission records for each survey participant and NDI records could be matched to any or all of the submission records created for a survey participant. Appendix A describes scenarios under which NCHS SPB generated alternate submission records.

Each record was screened to determine if it contained at least one of the following combinations of identifying data elements:

1. SSN (nine digits or last four digits), sex, full date of birth present
2. Last name, first name, month of birth, year of birth present
3. Last name, first name, SSN (nine or last four digits) present

Any survey participant submission records that did not meet these minimum data requirements were ineligible for record linkage. There were some records from the 1985 and 1997 NNHS where there was an indication of deceased status on the public-use files but participants were ineligible for NDI linkage. These cases are indicated as ineligible on the 2011 Linked Mortality Files.

Identifying potential match records

Potential NDI death record matches are first based on various combinations of matching identifiers between the files. Similar to previous linkages, NCHS SPB applied the seven criteria used by the NDI to identify potential matches. These included:

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 survey record matching birth surname on the NDI record (for females who change their name after marriage but don't supply a birth surname)

In addition to the seven criteria, NCHS SPB tested and included additional combinations of these identifiers to expand the number of potential matches.

Agreement on names was 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. The sound alike systems included both a variation of the New York State Identification Intelligence System (NYSIIS) and Soundex. NYSIIS converts a name to a phonetic coding. For example, records with last names Smith and Smyth received equivalent NYSIIS codes and both would be selected as a potential match for a NCHS submission with Smith (or Smyth) as a last name. Similarly, Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English. The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling, and is used in the same was as NYSIIS.

Any NDI record that matched an NCHS survey submission record on any one of the criteria was selected as a potential match. As one or more NDI records may have been matched to a given NCHS submission record, the NDI record selection process could return several potential matches for each person, many of which would be non-matches or duplicate records.

Scoring and classifying potential match records

As previously described, there are several ways that a NCHS survey participant submission record could match to an NDI record. For each potential match, a flag was created to indicate whether there was agreement or no basis for comparison for each identifier. NCHS SPB assigned a score to each potential match reflecting the degree of agreement between the identifying information on the NCHS survey submission record and the NDI death record. The score was based on probabilistic weights assigned to each of the identifying data items used in the NCHS -NDI record match. For example, a common first name, such as "John", that has a higher probability of occurrence in the population had a lower weight than an uncommon name such as "Jonas". Weights could be either positive or negative. If there was agreement between the NCHS survey record and the NDI record for a particular identifying data item, the weight was positive. If there was no agreement, the weight was negative. With the exception of middle initial, data items that were missing on the NCHS survey record, the NDI record, or both received a weight of zero. The score for each potential match was the sum of the weights for each individual data item.

$$\textbf{\textit{Score}}_{\text{ 9 digit SSN}} = \{W_{SSN1} + \ldots + W_{SSN9}\} + (p)W_{firstname\ x\ sex\ x\ birthyear} + W_{middleinitial\ x\ sex} + (q)W_{lastname} + W_{race} + W_{sex} + W_{maritalstatus\ x\ sex\ x\ age} + W_{birthdate} + W_{birthmonth} + (r)W_{birthyear} + W_{stateofbirth} + W_{stateof\ residence} + W_{survey\ middle\ /\ NDI\ first} + W_{survey\ middle\ initial\ /\ NDI\ first\ initial} + W_{surname\ |[female]}$$

W = weight
p,q,r = proration factors for inexact matches (including matches by NYSIIS/Soundex, initials, and year of birth +/- 3 years)
$SSNi$ = ith digit of the SSN

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

For NHIS surveys that only collected the last four digits of SSN, records were given the sum of the digit weights for the last four, as well as an additional digit (W_{SSNa}), making the assumption that if the last four digits matched, at least one of the first five digits matched. This gave records with four digit SSN's a slightly higher score for matching on the last four digits without giving an equivalent score to records with nine digit SSN's.

$$\textbf{\textit{Score}}_{\text{ 4 digit SSN}} = \{W_{SSNa} + W_{SSN6} + W_{SSN8} + W_{SSN7} + W_{SSN9}\} + (\text{p})W_{firstname\ x\ sex\ x\ birthyear} + W_{middleinitial\ x\ sex} + (\text{q})W_{lastname} + W_{race} + W_{sex} + W_{maritalstatus\ x\ sex\ x\ age} + W_{birthdate} + W_{birthmonth} + (\text{r})W_{birthyear} + W_{stateofbirth} + W_{stateof\ residence} + W_{survey\ middle\ /\ NDI\ first} + W_{survey\ middle\ initial\ /\ NDI\ first\ initial} + W_{surname\ |[female]}$$

After scoring the potential matches, each was categorized into one of five mutually exclusive classes. Whereas weighting and scoring take into account the probability that the NCHS survey 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). The classes do not necessarily use all of the individual data items used in creating the score.

As SSN is a key identifier in the matching process, each NCHS-NDI record match was initially classified according to whether a 4 or 9 digit SSN was present and agrees (Class 1 or 2), was present but disagrees (Class 5) or was missing (Class 3 or 4). The five classes used by SPB for the NCHS 2011 Linked Mortality file were as follows.

Class 1: Agreed on at least 8 (of 9) or 4 (of 4) digits of SSN, first name (including NYSIIS/Soundex match), middle initial (including blank), last name (including NYSIIS/Soundex match), birth year (+/- 3 years), birth month, sex, and state of birth.

Class 2: Agreed on at least 7 (of 9) or 4 (of 4) digits of SSN at least 5 more of the following items: first name (including NYSIIS/Soundex match), middle initial (including blank), last name (including NYSIIS/Soundex match), birth year (+/- 3 years), birth month, sex, and state of birth.

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

Type A: SSN is unknown, but last name matched (including NYSIIS/Soundex match) and at least 7 of the following items agreed: first name (including NYSIIS/Soundex 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 if after review, there was the possibility that SSN was either recorded incorrectly or that the spouse's SSN was recorded instead of the subject's SSN. In this category, SSN was known but 3 or more (of 9) and 1 or more (of 4) digits did not agree, but at least 8 of the following items agreed: first name, middle initial (including blank), last name, birth year, birth day, sex, race, marital status, and state of birth. 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 was unknown on either the NCHS survey submission record or the NDI record and fewer than 8 of the items listed in Class 3 matched.

Class 5: SSN was present but fewer than 7 (of 9) or 4 (of 4) digits on SSN agreed.

Selecting matches and assigning vital status

Since each eligible NCHS survey participant may have had multiple submission records and each submission record may have returned one or more potential matches to an NDI record, NCHS SPB employed a strategy to provide the single best NDI match record for inclusion on the linked mortality file. First, NCHS-NDI potential match records that had a date of death prior to the date of interview or a score of zero or less were considered false matches and were 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, and different records could potentially end up in different classes. The remaining potential matches were ranked first on class (from 1 to 4) and then within class by highest score. NCHS SPB selected the NDI match with the highest score within the best class (if in class 1 or 2) or the highest score only (if in class 3 or 4). In the event of a tie among NDI record matches for a particular NCHS survey record, the record underwent manual review with the tiebreaker reflecting the importance of matching items.

Next, NCHS SPB 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. 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 44, 45, and 42, 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 NCHS survey participants that fell below the cut-offs, NCHS SPB conducted a manual review of the NCHS survey submission record and the corresponding NDI potential matches to determine vital status.

Appendix A

Creating Alternate Submission Records

The primary purpose of using alternate submission records was to increase the chances of returning a correct death record for those survey participants who were, in fact, deceased. Alternate submission records may be created for several reasons. For example, if an SSN was present but additional information indicates that the SSN was not valid, an alternate submission record will be created that does not include the SSN. Name inaccuracies were 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 were created that took into account nicknames being listed as the first name, using a nickname to proper name conversion process or that used 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 had both an imputed month of birth (12 separate records) and two-part first name (3 separate records) resulting in 36 NDI submission records.