The Third National Health and Nutrition Examination Survey (NHANES III)
Linked Mortality File: Matching Methodology
December 2005

Introduction

The NHANES III Linked Mortality file represents the first restricted release of linked NHANES III records to the National Death Index (NDI). This file links NHANES III participants, age 17 and older, with death records from the (NDI) through December 31, 2000. The NDI is a central computerized database of all certified deaths in the United States since 1979. For detailed information on the NDI's contents and methods, refer to the National Death Index (NDI). NHANES III (1988-94) was designed to provide national estimates of the health and nutritional status of the civilian noninstitutionalized population of the United States aged 2 months and older. Details of the survey design and questionnaires are published in the NHANES III Plan and Operation reference manual (NCHS 1994, U.S. DHHS 1996). This new resource allows researchers the opportunity to conduct a vast array of outcome studies designed to investigate the association of a wide variety of health and risk factors with mortality.

Overview

The linking of NHANES III and NDI records was conducted by probabilistic matching. NCHS employed a matching methodology for the NHANES III Linked Mortality file that is similar, but not identical, to the standard methodology offered by the NDI. Specifically, NCHS developed new weights associated with the specific value of each identifying element on the submission record to create scores for potential matches and implemented more restrictive criteria for identifying potential matches than the NDI standard approach (see Step 3 below). Also, NCHS conducted a new calibration study to establish the cut-off scores for determining whether a NDI match is considered a true match or a false match (Step 4 below). Additionally, for a selected sample of NHANES III records, NCHS reviewed death certificates to verify whether the NHANES III and NDI record match was correct.

This document explains the matching methodology NCHS employed to link NHANES III records to death records in the NDI. In this document, data users will find detailed information on the following steps involved in the NHANES III-NDI match process.

1. Creating NDI submission records from NHANES III respondent records
2. Selecting potential matches between NHANES III and NDI records
   - Selection is based upon 7 different criteria
   - Selection creates a pool of potential matches
3. Scoring and classifying potential matches
   - Scores are based upon weights associated with the *values* of each identifying data item
   - Classes are based upon *which* identifying data items match
4. Determining final match status and assigning vital status

Figure 1 depicts the NDI matching process for NHANES III. Users interested in a detailed description of the standard NDI matching methodology should refer to the National Death Index (NDI).

NHANES III-NDI record linkage

1. Submission Records

Only NHANES III participants 17 years of age or older at the time of interview, with sufficient identifying information, were eligible for matching with NDI records.[1] For each eligible NHANES III participant, NCHS prepared a base submission record for the NDI that contained up to 12 identifying data items (see below). The NHANES III routinely collects all of the 12 data items used by the NDI for matching and has essentially 100 percent complete reporting of these items except for social security number (SSN) and middle initial (see table 2 in Tabular data). Names collected during the NHANES III interview such as "John Doe", "Female #1", and "Person" are deleted as non-valid names. Names containing obvious keystroke errors such as "Jo9hn" are corrected.

Data items on the NHANES III submission record

1. Social Security Number
2. First name
3. Middle initial
4. Last name (or birth surname)
5. Month of birth
6. Day of birth
7. Year of birth
8. Sex
9. State of birth
10. Race
11. State of residence
12. Marital status

In addition to the base submission record, the NDI allows multiple alternate submission records. In order to increase the chances for selection of the correct death record, NCHS generated alternate submission records, e.g. when identification data were questionable or when the NHANES III participant had a multi-part name. For a detailed description of the rules NCHS used to generate alternate NHANES III submission records, refer to Appendix A.

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

Any submission record that did not meet these minimum data requirements was ineligible for record linkage. Table 1 (Tabular Data) provides a description of the number of survey participants and eligibility status. All accepted NHANES III submission records are further edited by the NDI system to provide a consistent format for identifying data elements before

[1]A very small number of adult NHANES III participants (N = 26) were ineligible for mortality linkage because they lacked sufficient identifying information to create a submission record.

performing the NDI record search and retrieval process. For example, the NDI editing process converts text to all upper case and removes suffixes from last names. Also, since spelling variants of names are common, NDI codes last names based on the way a name sounds rather than how it is spelled[2]. For example, records with last names Smith and Smyth receive equivalent NYSIIS codes and both would be selected as a potential match for a NHANES III submission with Smith (or Smyth) as a last name.

2. Selecting NHANES III-NDI potential match records

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 NHANES III-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)

Any NDI record that matches a NHANES III submission record on any one of these seven criteria is selected. As one or more NDI records may be matched to a given NHANES III submission record, the NDI record selection process can return several hundred *potential* matches for each NHANES III person, many of which will be non-matches or duplicate records.

3. Scoring and classifying potential match records

Assessing the quality of potential matches and determining the best match for each NHANES III participant requires a consistent approach. The matching methodology begins by assigning probabilistic scores for each potential match. The score is the sum of the weights assigned to each of the identifying data items used in the NHANES III-NDI record match, where the weights reflect the degree of agreement between the information on the NHANES III submission record and the NDI death record. 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 $[\text{Log}_2 (1/p_i)]$: the base 2 logarithm of the inverse of the probability of occurrence of the value of the identifying data item on the submission record. Two examples of how weights are created for identifying data items are as follows:

- Social Security number – each digit in each position of the SSN (1 to 9) has a corresponding binit weight, with the total SSN weight being the sum of the weights for each of the nine digits. For a record to receive the total SSN binit weight at least 8 digits need to agree. 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.

[2] 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.

- Name – a common first name, such as "John", that has a higher probability of occurrence has a lower weight than an uncommon name such as "Jonas". First name weights are stratified by both sex and year of birth since first names are sex specific and the popularity of first names varies over time. Weights for first and last names are limited to a finite set of values and any name not appearing in the set (meaning it is less common) receives the maximum value of the weight.

Weights are either positive or negative. If there is agreement between the NHANES III record and the NDI record for a particular identifying data item, the weight is positive. If there is no agreement, the weight is negative. Some items, such as year of birth, allow a tolerance (+/- 3 years) and are still considered to agree. With the exception of middle initial, data items that are missing on the NHANES III submission record, the NDI record, or both receive a weight of zero. A blank middle initial is considered a valid value and receives the appropriate weight. The score for each potential match is the sum of the weights for each individual data item.

$$\textbf{\textit{Score}} = \{\sum W_{SSN1} + \ldots + W_{SSN9}{}^{3}\} + W_{firstname\ x\ sex\ x\ birthyear} + W_{middleinitial\ x\ sex} + W_{lastname} + W_{race} + W_{sex} + W_{maritalstatus\ x\ sex\ x\ age} + W_{birthdate} + W_{birthmonth} + W_{birthyear} + W_{stateofbirth} + W_{stateof\ residence}$$

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 NHANES III 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. For example, as SSN is a key identifier in the matching process, each NHANES III-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). Additionally, non-changing identifying information is more important than information that can change over time. A common example of a legitimate change in information over time is when a woman assumes her spouse's surname at marriage. Birth surname, however, does not change and is thus an important matching variable for women. By contrast, state of residence and marital status may change between the NHANES III interview date and the date of death and are, therefore, less important as matching variables.

The final five classes used by NCHS for the 2004 NHANES III 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.

[3] 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.

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 month, 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[4]. SSN is known but 3 or more digits do not agree, but at least 8 of the following items agree: first name (including NYSIIS match), middle initial (including blank), last name (including NYSIIS match), birth year (+/- 3 years), birth month, sex, race, marital status, and state of birth. Last name and sex must agree.

Class 4: SSN is unknown on either the NHANES III 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 or at least 7 digits on SSN agree but fewer than 5 of the following items agree: 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.

4: Selecting matches and assigning vital status

As already described in section 2, each eligible NHANES III participant may have multiple submission records and each submission record may return one or more potential matches to a NDI record. The NHANES III Linked Mortality File does NOT include all the potential matching NDI records. Rather for those NHANES III participants with a potential match to the NDI, NCHS employed a strategy to provide the single best NDI match record for inclusion on the linked mortality file.

First, NHANES III-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, for each eligible NHANES III participant. The NDI match with the highest score within the best class was selected as the single best record match. In the event of a

[4] 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.

tie among NDI record matches for a particular NHANES III record, the tiebreaker reflected the importance of matching items[5].

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 Study (NHEFS) calibration sample, which has verified mortality outcomes for its 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. Users should refer to Appendix B for a description of the results of the calibration study.

In addition, NCHS reviewed and verified death certificates for a select sample[6] of the 5,630 NHANES III participants, who had a single best NDI record match chosen. In the death certificate review process, NCHS examined information, such as industry and occupation, spouse's name, and mail address, in addition to the NDI's nine matching criteria, to determine whether a potential match correctly matched the NHANES III participant. NCHS reviewed a total of 2,544 death certificates, from Classes 1 through 4, of which 2,521 were considered "true" matches by the probabilistic matching process (assumed deceased) and 23 were considered "false" matches (assumed alive). Among the 2,521 assigned decedents, 98.8% were confirmed deceased after death certificate review. Among the 23 persons assumed to be alive for whom we reviewed the death certificate for the best matching NDI record, 3 were found to be deceased. The final mortality status assignment is 3,384 NHANES participants assumed deceased. Figure 2 shows the breakdown of NDI record matches receiving a death certificate review, the determination of "true" and "false" matches, and the match status confirmed by death certificate review.

[5] The order is: number of digits of SSN; sex; last name; first name; state of birth; year of birth; month of birth; day of birth. If all of these are the same, then a random number is used.

[6] The sample of death certificates reviewed was not randomly selected. Thus, the results of the death certificate review are not intended to be used as a confirmatory study of the probabilistic matching process. See Appendix B for a description of the calibration study.

Notice to Users

Eligible NHANES III participants with a "true" NDI record match are assumed to be dead. Eligible NHANES III participants with no potential NDI record match as well as those with a NDI record match that is considered a "false" match are assumed to be alive. Ineligible NHANES III respondents should be excluded from mortality analyses. Analysts should use the variable MORTSTAT to determine vital status.

A data file with the additional probabilistic NDI match results is available by request. This special request file differs from the current file in that not every NHANES III participant with a NDI record is considered deceased. The special request file includes NDI record match results for potential NDI matches that were considered "false" by the probabilistic matching algorithm. NCHS has provided the SCORE and CLASS for the best NDI record match, regardless of the final assigned vital status, to provide the user with the opportunity to alter the criteria for determining final match status. The user can take either a more or less conservative approach to vital status ascertainment by setting a different cut-off score within each class and/or determining which classes contain true matches. For more information on the implications of using alternate cut-off scores on vital status ascertainment, please see Appendix C in the NHIS matching methodology report.

Figure 1: NHANES III-NDI Matching Process

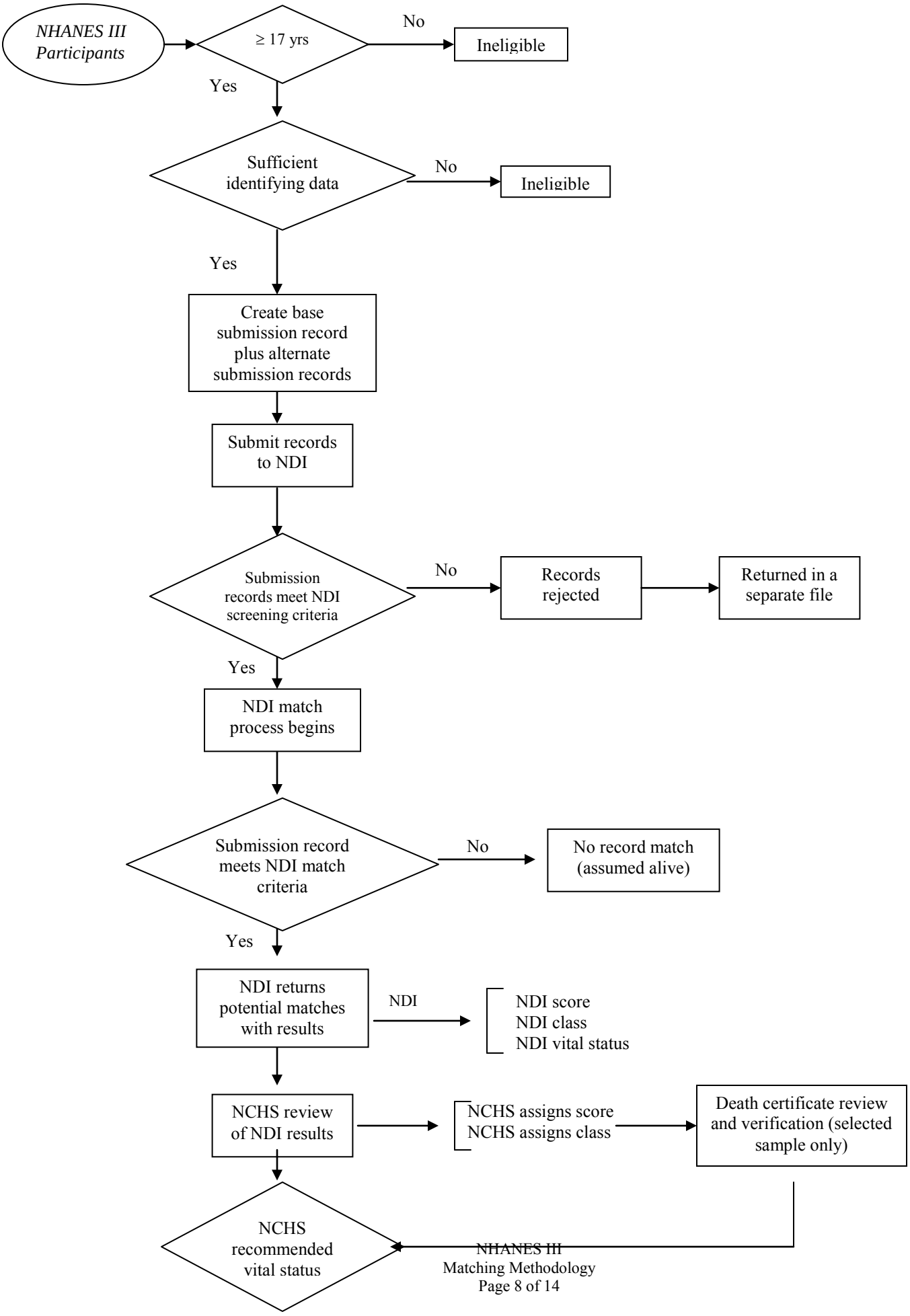
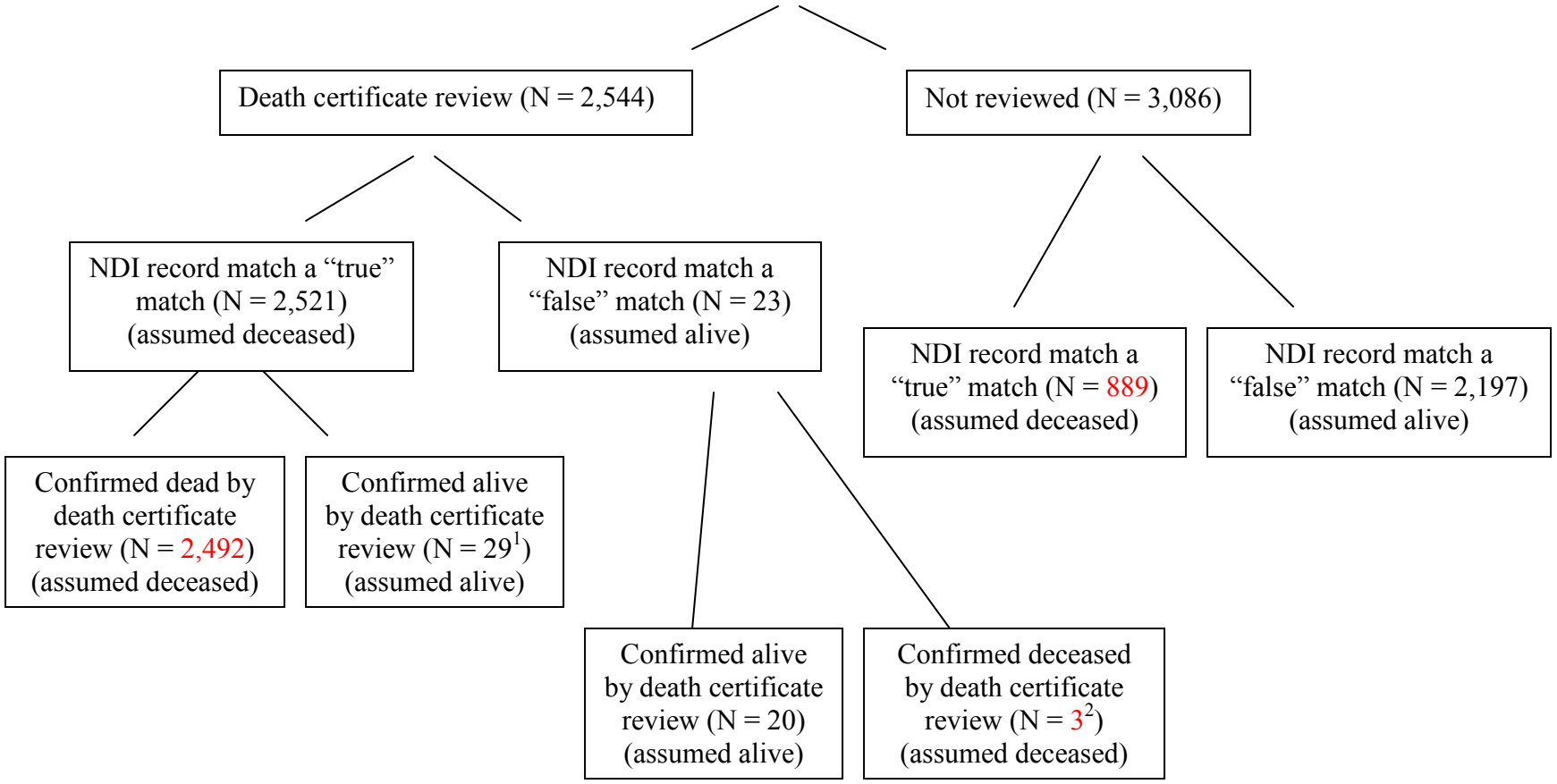


Figure 2.
NHANES III
Single best NDI record match
(N = 5,630)



[1]These 29 cases had their status changed to alive.
[2]These 3 cases had their status changed to deceased
Note: The total number of NHANES decedents = (2,492 + 3 + 889) = 3,384

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 NHANES III participants who are, in fact, deceased. The NDI allows multiple alternate submission records for each survey person. Rules for creating alternate NDI submission records were based upon a calibration study using the NHANES I Epidemiologic Follow-up Study (NHEFS). The NHEFS calibration study has a sample of 12,699 people whose vital status is known for a definite time period beginning January 1979 through either the date of death for decedents or a final interview date for non-decedents. NCHS created base submission records for this sample and submitted them to the NDI record retrieval process. For those known to be deceased but who did not return an NDI record match, NCHS compared the identifying information on the submission record to the information on the death certificate. The process revealed the most common reasons a NDI record was not returned.

Name inaccuracies are the most common type of mismatch error encountered when matching to the NDI system. Since death certificates are official records, they will list the full proper name of the decedent. However, survey respondents may provide nicknames or middle names as their first names. To account for nicknames being listed as the first name, NCHS used a nickname to proper name conversion process that created alternate submission records with the most popular formal name associated with that nickname. For example, if a NHANES III record listed the respondent name as Beth, two submission records were created. The base submission record included Beth as the first name and the alternate submission record included Elizabeth as the first name.

Multipart first or last names also increase the chances of a NHANES III and NDI record not matching. Such differences in name reporting are particularly common for the U.S. Hispanic population. For example, mother's and/or father's surname may both be reported as two last names in a particular order during the survey contact but may be reversed on the death record. To take into account potential recording discrepancies caused by multi-part names, alternate records were created using all of the components of multi-part names both separately and together. Only names with either a space or hyphen are treated as multipart names. Middle initial plays an important role in NDI matching. Since the NDI allows a blank as a valid value for middle initial, an alternate record is created by dropping the middle initial from any base submission record where it is non-blank.[1]

In summary, for the NHANES III-NDI linkage, the following rules were used for generating alternate submission records:

1. Use proper name in place of nickname for first name
2. Multipart first and last names are submitted as is, and alternately each part of the name is submitted as the first or last name
3. Switch first name and middle name

[1] Preliminary research performed at NCHS has found that many survey data files include a blank middle initial about 25% of the time, making blank the single most commonly reported middle initial.

4. Blank out middle name
5. Add alternate surnames when evidence of a legal name change is available
6. Use alternate birth date or SSN data, if collected
7. If month of birth is missing, submit twelve records, one with each month

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

NHANES I Follow-up Calibration Sample

Since the NDI record selection and match processes do not have an independent means of assessing whether a NHIS-NDI match is true or false, NCHS undertook a calibration study to determine the adequacy of the probabilistic approach utilized to match NHANES III participants to NDI records. Such a study is necessary in order to assess the number of false negatives and false positives.

With regard to false negatives, there are several ways the death of a NHANES III participant could be missed. Some of these ways are due to the universe of deaths in the NDI, some to the NDI selection process and some to the ranking, scoring and classification of matches employed by NCHS in the NHANES III-NDI linkage (see sections 3 and 4 of the main document). Specifically, there are five ways a death could be missed in the NHANES III mortality files:

- Deaths outside the United States are not included in the NDI database;
- A small number of deaths occurring in the U.S. are not part of the NDI database;
- Deaths not retrieved in the NDI selection process;
- True deaths retrieved in the NDI record selection process are dropped from the pool of potential matches because they are not the top ranked death record by NCHS;
- True deaths retrieved in the NDI record selection process are assigned a score below the threshold for determining a match a true match.

False positives often arise by finding a match for a relative or someone with a common name. A small number of false positives also occur when true decedents are matched to the wrong NDI record. Although these individuals are assigned the correct vital status, as the death record is wrong, the date and cause of death are unlikely to be correct.

The calibration study used the NHANES I Follow-up survey (also known as NHEFS), which was conducted from 1971-1975. NHEFS provides a unique opportunity to assess the quality of the NHANES III-NDI matching process because it is a longitudinal study with a high participation rate and highly complete and verified identification data. In the NHEFS sample, there are 12,699 people for whom active follow-up was conducted so that their vital status was known beginning January 1979[1] through either the date of death or a final interview date (for non-decedents). In this sample, four deaths occurred outside the United States, leaving 3,454 deaths that were available to be included in the NDI database and for which a match to a NHEFS participant was possible. NCHS applied the same approach for creating submission records, selecting NDI records, and ranking, scoring, and classifying matches to the NHEFS sample as for the NHANES III-NDI linkage to determine how many of the 3,454 deaths could be found[2].

Figure 3 depicts the selection process and match status determination of the NHEFS sample. Among the 3,454 NHEFS decedents, 3,380 had a NDI record selected as a potential match and

[1] The NDI was established in 1979. Persons in the NHEFS sample who died before 1979 were not considered in this study.

[2] As noted, a small number of deaths that occur in the U.S. are missed by the NDI database. In this case, two NHEFS deaths that occurred in the U.S. were not included in the NDI database.

74 did not. Among the NDI potential matches for the NHEFS decedents, 3,375 had the *correct* NDI record selected. Using the cut-off scores for Classes 2, 3, and 4 as described in section 4 of the documentation, resulted in 3,322 being considered true matches and correctly assigned as deceased, whereas 53 were considered false matches and incorrectly assigned a vital status as alive. Overall, there were 79 NHEFS decedents who did not have a true match to a NDI record - 5 were decedents who were assigned as dead, but because their NDI record match is to the wrong person, the date and cause of death will not be correct (a form of false positive) and 74 were decedents who did not return a NDI record and were incorrectly assigned a vital status of alive. Among NHEFS non-decedents, 49 returned a NDI record that was selected as a true match and were incorrectly assigned a vital status of deceased.

Table 1 shows the cut-off scores for Classes 2, 3, and 4 employed to determine the match status of NDI potential matches to NHEFS records and the proportion correctly classified. Based upon the matching methodology described for the NHIS Linked Mortality files, across all four classes, 96.1% of NHEFS decedents were correctly classified as deceased and matched to the correct death certificate, 99.4% of non-decedents were correctly classified as alive, with an overall 98.5% of NHEFS respondents correctly classified.

Table 1:Cut-off scores and proportion of NHEFS subjects correctly classified.

Within Class	Cut-off Score	Correctly classified overall (%)	Correctly classified as dead (%)	Correctly classified as alive (%)
2	≥ 47	98.0	98.5	40.0
3	≥ 45	89.7	94.7	67.5
4	≥ 40	98.6	60.5	99.4
Total across classes	**-----**	**98.5**	**96.1**	**99.4**

Figure 3. NHEFS Calibration Study