

The NHANES I Epidemiologic Follow-up Study (NHEFS) Linked Mortality File: Matching Methodology December 2005

Introduction

The 2005 restricted data release of the NHANES I Epidemiologic Follow-up Study (NHEFS) represents a mortality update for the NHEFS cohort. This current file includes death information for deaths ascertained during the NHEFS follow-up periods in 1982-84, 1986, 1987, and 1992 as well as new record linkages of NHEFS participants to death records from the National Death Index (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 <http://www.cdc.gov/nchs/ndi.htm>. The NHEFS uses as its baseline the 14,407 adult persons (25-74 years), who completed a medical examination between 1971 to 1975 as part of the first National Health and Nutrition Examination Survey (NHANES I). Details of the survey design and follow-up periods are published in the Plan and Operation of the [NHANES I Epidemiologic Follow up Study](#), 1992 (Cox et al., 1997). 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 collected over time with subsequent mortality.

At the conclusion of the NHEFS data collection period, there were 4,604 NHEFS participants identified as deceased, 9,251 participants successfully traced alive during at least one follow-up data collection wave, and 546 participants who were never successfully traced and were considered lost to follow-up for the entire NHEFS study period (1982 through 1992). During the original NHEFS mortality follow-up, a participant was not considered deceased unless a proxy respondent completed a proxy decedent interview or a death certificate was collected and verified. NCHS submitted records for all 14,407 participants to the NDI. For the 4,604 confirmed NHEFS decedents, this match to NDI records provided a way to (1) confirm the probabilistic matching process and (2) obtain cause of death information if it was not available from the proxy decedent interview. For the NHEFS participants who were known to be alive as of 1992, the match to NDI records provided the opportunity to search for deaths that had occurred in the longitudinal cohort since 1992. Finally, for those NHEFS participants who were unable to be traced during one of the follow-up periods, the match to NDI records provided the opportunity to search for deaths that were missed due to loss to follow-up.

Overview

The updated NHEFS mortality file includes deaths previously ascertained during one of the follow-up periods plus the additional mortality follow-up obtained by linking NHEFS and NDI records. The NHEFS-NDI record linkage was conducted by probabilistic matching. NCHS employed a methodology for the NHEFS Linked Mortality file that was 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 including potential matches than the standard NDI 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

NHEFS records, NCHS reviewed death certificates to verify whether the NHEFS and NDI record match was correct.

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

1. Creating NDI submission records from NHEFS respondent records
2. Selecting potential matches between NHEFS 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 NHEFS. Users interested in a detailed description of the standard NDI matching methodology should refer to the [NDI](#).

NHEFS-NDI record linkage

1. Submission Records

For each NHEFS participant, NCHS prepared a base submission record for the NDI that contained up to 12 identifying data items (see below). The NHEFS collected 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 ([Tabular data](#)).

Data items on the NHEFS 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 NHEFS participant had a multi-part name. For a detailed description of the rules NCHS used to generate alternate NHEFS 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

Submission records that do not meet these minimum data requirements are ineligible for record linkage by the NDI.

There were no NHEFS participants who were ineligible because they did not meet the minimum NDI data requirements. All accepted NHEFS submission records are further edited by the NDI system to provide a consistent format for identifying data elements before employing 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 spelled¹. For example, records with last names Smith and Smyth receive equivalent NYSIIS codes and both would be selected as a potential match for a NHEFS submission with Smith (Smyth) as a last name.

2. Selecting NHEFS-NDI potential match records

The [NDI](#) system selects death record matches based on a set of established match criteria. The seven criteria listed below are the criteria in use at the time of the current NHEFS-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 NHEFS submission record on any one of these seven criteria is selected. As many NDI records may be matched to a given NHEFS record, the NDI record selection process can return several hundred *potential* matches for each NHEFS participant, 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 NHEFS participant requires a consistent approach. The matching methodology begins by assigning

¹ 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.

probabilistic scores for each potential match. The score is the sum of a set of weights assigned to each of the identifying data items used in the NHEFS-NDI record match, where the weights reflect the degree of agreement between the information on the NHEFS 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 be assigned 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.
- Name – a common 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 NHEFS 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 NHEFS 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.

$$\text{Score} = \{ \sum W_{SSN1} + \dots + W_{SSN9}^2 \} + W_{\text{firstname} \times \text{sex} \times \text{birthyear}} + W_{\text{middleinitial} \times \text{sex}} + W_{\text{lastname}} + W_{\text{race}} + W_{\text{sex}} + W_{\text{maritalstatus} \times \text{sex} \times \text{age}} + W_{\text{birthdate}} + W_{\text{birthmonth}} + W_{\text{birthyear}} + W_{\text{stateofbirth}} + W_{\text{stateofresidence}}$$

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 NHEFS 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 NHEFS-NDI record match is initially categorized into 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

² 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.

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 thus is an important matching variable for women. State of residence and marital status may change between the NHEFS interview date and the date of death and are less important as matching variables.

The final five Classes used by NCHS for the NHEFS 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 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 3³. 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 NHEFS 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 NHEFS participant may have multiple submission records and each submission record may return one or more potential matches to a NDI record.

³ 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

The NHEFS Linked mortality file does NOT include all of the potential matching NDI records. Rather for those NHEFS 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, NHEFS-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, duplicate match records (i.e. those 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 NHEFS participant. The NDI match with the highest score within the best class was selected as the single best record match. In the event that there was a tie among NDI record matches for a particular NHEFS record, the tiebreaker reflected the importance of matching items⁴.

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 within Classes 2, 3, and 4 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.* These scores were chosen because they 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.

The NHEFS – NDI Linked Mortality Data File includes 6,656 NHEFS decedents – 2,052 deaths identified through NDI record matches plus the 4,604 NHEFS participants who were identified as deceased at the completion of the final NHEFS data collection period. NHEFS participants were identified as deceased by one or more mortality data sources: proxy decedent interview, death certificate collection, and/or probabilistic NDI record matches. In all but three cases, the previously collected death information for the 4,604 deaths was confirmed. In those three instances, higher scoring NDI death records were found and the previously collected death certificate data was replaced with the corrected date and cause of death data.

Of the 6,656 NHEFS deaths, 5,715 (86%) were verified by death certificate review. 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 NHEFS participant. NCHS could not verify 941 deaths with a death certificate. Among these, 75 were identified by proxy interview only and 866 were determined to be deceased according to the NDI probabilistic scoring criteria. Ninety-five percent of the deaths identified only through probabilistic matches were classified as excellent or very strong matches (Class 1 or Class 2) indicating a high likelihood that the death information is correct.

⁴ 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.

Vital Status Ascertainment of NHEFS participants as of 1992 Follow-up and the current Mortality Update

	NHEFS Mortality Update		
NHEFS 1992 Follow-up	Alive	Deceased	Total
Alive	7,253	2,004	9,257
Deceased	0	4,604	4,604
Lost to Follow-Up	498	48	546
Total	7,751	6,656	14,407

Notice to Users

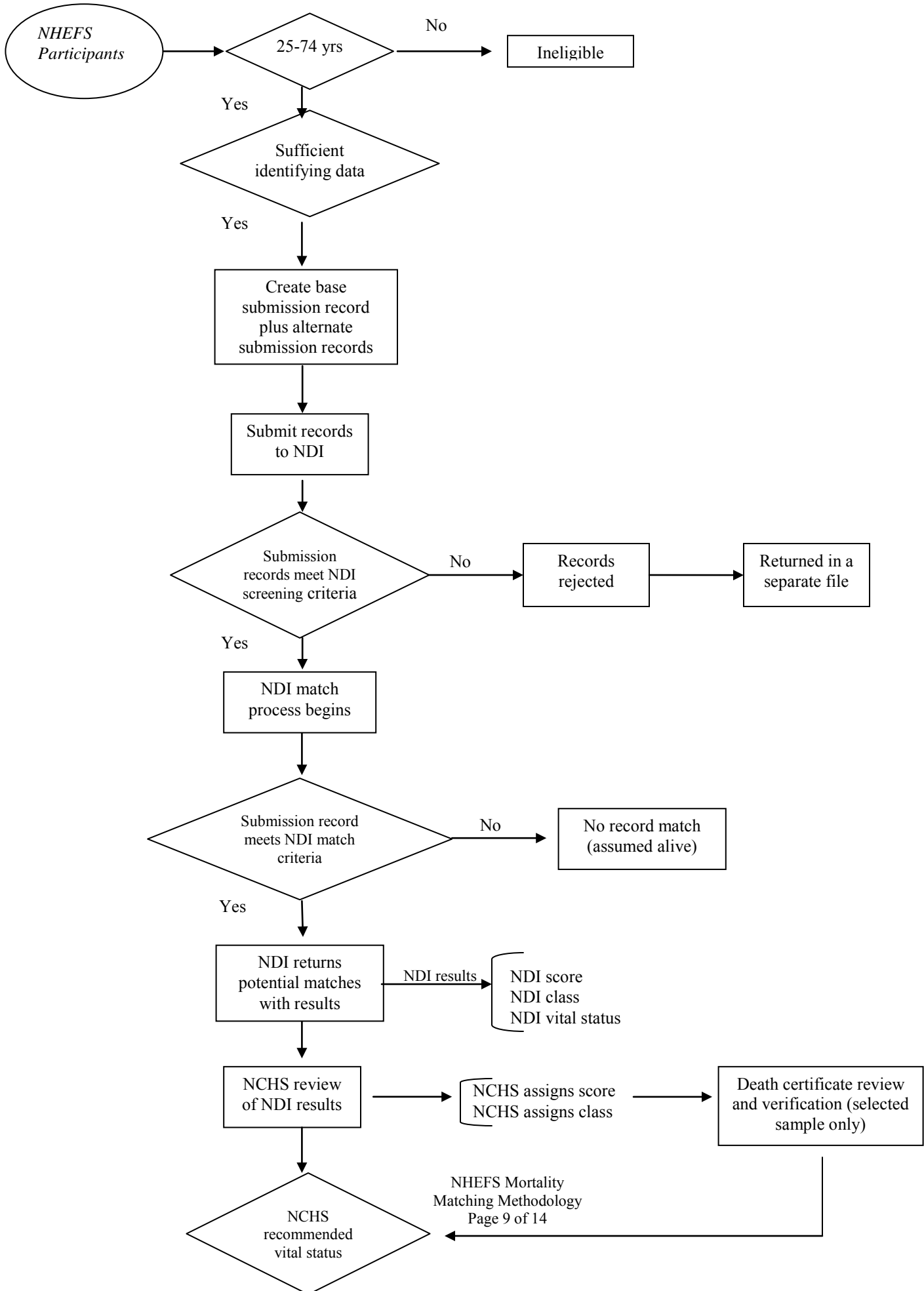
NHEFS participants with a “true” NDI record match are assumed to be dead, while those with no potential NDI record match as well as those with a NDI record match that is considered a “false” match are assumed to be alive. Analysts should use the variable MORTSTAT to determine vital status.

For the 546 NHEFS respondents who were never located or interviewed in any of the four NHEFS data collection periods, 48 were found to be deceased in the current probabilistic NDI match. The remaining 498 lost to follow-up respondents are assumed to be alive based on the criteria used for assigning vital status through a probabilistic NDI record match. However, some researchers may wish to take a more conservative approach and consider the remaining 498 NHEFS lost to follow up participants as having an unknown vital status and exclude them from mortality analyses. The variable MORSTAT indicates the assigned vital status of the NHEFS participant. In addition, the variable VITST92 has been provided on the NHEFS standard mortality file to identify the NHEFS participants who have never been successfully re-contacted so that researchers can identify these 498 individuals.

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 NHEFS 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](#).

Since the linkage of NCHS survey records to the NDI is entirely probabilistic, NCHS conducted a calibration study to establish the cut-off scores that are used to determine whether a NDI match is considered true or false. Such a study requires applying the matching methodology described in this report to a population that has verified mortality outcomes. A portion of the NHEFS sample (N=12,699) with confirmed vital status was used to validate the NDI matching methodology. Users should refer to [Appendix B](#) for a description of the results of the calibration study.

Figure 1: NHEFS-NDI Matching Process



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 NHEFS 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 NHEFS 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 NHEFS 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 NHEFS-NDI linkage, the following rules were used for generating alternate submission records:

1. Use proper name in place of nickname for first 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.

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
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

NCHS undertook a calibration study to determine the adequacy of the probabilistic approach utilized to match NCHS survey 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 that a death to a survey 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 (see sections 3 and 4 of the main document). Specifically, there are five ways a survey participant's death could be missed:

- 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 are not retrieved in the NDI record 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 1971-1992. NHEFS provides a unique opportunity to assess the quality of the 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 used for the calibration study, there were 12,699 people for whom active follow-up was conducted so that their vital status was known beginning January 1979¹ 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 approach to linking with NDI that includes creating submission records, selecting NDI records, and ranking, scoring, and classifying matches to the NHEFS sample to determine how many of the 3,454 deaths could be found².

Figure 2 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 74 did not. Among the NDI potential matches for the NHEFS decedents, 3,375 had the *correct*

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.

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 this matching methodology, 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 2. NHEFS Calibration Study

