

**The Second National Health and Nutrition Examination Survey (NHANES II)
Linked Mortality File: Matching Methodology
December 2005**

Introduction

The 2005 restricted release of the Second National Health and Nutrition Examination Survey (NHANES II) Linked Mortality File represents a mortality update for the 1992 NHANES II Mortality Study. This current file includes death information for deaths ascertained as part of the 1992 mortality follow-up as well as new record linkages of adult NHANES II participants with 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 the [NDI](#). This new resource allows researchers the opportunity to conduct a vast array of outcome studies designed to investigate the associations of a wide variety of health and risk factors with mortality.

The prior mortality linkage included a subset of NHANES II participants, aged 30 years and older and who agreed to a physical examination (N = 9,252). The study ascertained vital status of its participants through December 31, 1992 by two methods: (1) for deaths occurring between 1979 and 1992, the study linked records with the NDI and (2) for deaths prior to 1979, it linked records to the Social Security Administration's Death Master File (DMF). At the conclusion of the mortality follow-up, there were 2,145 participants who were assumed to be deceased. For the updated file, NCHS created submission records for all 9,252 participants of the original study. The current NHANES II mortality linkage updates the previous linkage in several ways. First, the updated linkage of the NHANES II to the NDI provides the opportunity to search for deaths that have occurred since December 31, 1992. Second, an enhanced matching process allowed NCHS to augment previous decedent information, when necessary. Finally, the majority of NHANES II participants who were identified as deceased through a NDI match were confirmed as deceased by death certificate review.

Overview

The updated NHANES II mortality file was conducted primarily by probabilistic matching to the NDI. For deaths prior to 1979, the previously ascertained vital status from the Social Security's Death Master File was retained. For the current NHANES II-NDI record linkage, NCHS employed a matching methodology for the NHANES II 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 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 the majority of NHANES II-NDI matches, NCHS collected death certificates to determine whether the NHANES II and NDI record match was correct.

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

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

NHANES II-NDI record linkage

1. Submission Records

For each eligible adult included in the 1992 NHANES II Mortality Study, NCHS prepared a base submission record for the NDI that contained up to 10 identifying data items (see below)^{1,2}. For the 10 data items, with the exception of middle initial, NHANES II has essentially 100 percent complete reporting of these items ([Tabular data, Table 2](#)).

Data items on the NHANES II submission record

1. First name
2. Middle initial
3. Last name (or birth surname)
4. Month of birth
5. Day of birth
6. Year of birth
7. Sex
8. State of birth
9. Race
10. 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

¹ A very small number of adult NHANES II participants (N=3) were ineligible for mortality linkage because they lacked sufficient identifying information to create a submission record.

² NHANES II did not collect Social Security numbers (SSN) or state of residence. Standard NDI submission records contain up to 12 identifying data items because of the inclusion these identifiers.

when the NHANES II participant had a multi-part name. This was a new process not included in the previous NHANES II mortality study. 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

Since SSN was not collected for NHANES II participants, NHANES II submission records had to meet screening criteria number 2. All accepted NHANES II submission records were 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 spelled³. 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 II submission with Smith (or Smyth) as a last name.

2. Selecting NHANES II-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 current NHANES II-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)

Since SSN was not collected for NHANES II participants, any NDI record that matches a NHANES II submission record on any one of the criteria two to seven is selected. As many NDI records may be matched to a given NHANES II record, the NDI record selection process can return several hundred *potential* matches for each NHANES II participant, many of which will be non-matches or duplicate records.

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.

3. Scoring and classifying potential match records

Assessing the quality of potential matches and determining the best match for each NHANES II 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 II-NDI record match, where the weights reflect the degree of agreement between the information on the NHANES II submission record and the NDI death record. NCHS developed the weights, known as binit weights, based upon the frequency of occurrence of the 10 data items in the NDI files for years 1979 to 2000, which represents about 49 million persons. For example, a common name, such as “John”, that has a higher probability of occurrence has a lower binit weight than an uncommon name, such as “Jonas”. 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.

Weights are either positive or negative. If there is agreement between the NHANES II 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 II submission record, the NDI record, or both receive a weight of zero. For all NHANES II records, the weights for SSN are 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.

$$\text{Score} = \{ \sum W_{SSNI} + \dots + W_{SSN9}^4 \} + W_{\text{firstname} \times \text{sex} \times \text{birthyear}} + W_{\text{middleinitial} \times \text{sex}} + W_{\text{lastname}} + W_{\text{race}} + W_{\text{sex}} + W_{\text{maritalstatus} \times \text{sex} \times \text{age}} + W_{\text{birthdate}} + W_{\text{birthmonth}} + W_{\text{birthyear}} + W_{\text{stateofbirth}} + W_{\text{stateofresidence}}$$

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 NHANES II 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 NDI identifying items are more important for determining true matches than others. For example, SSN is a key identifier in the NDI matching process and each 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). Because SSN is missing for NHANES II records, all matches were 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, marital status

⁴ 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.

may change between the NHANES II interview date and the date of death and is, therefore, less important as a matching variable.

The five Classes used by NCHS for NDI matches 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, and state of birth.

Type B: Records in this category were initially put in Class 5 but switched to Class 3⁵. 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 II 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.

As SSN was not collected for the NHANES II-NDI record linkage, all NHANES II-NDI potential matches only could be Class 3 or 4 matches.

4: Selecting matches and assigning vital status

As already described in section 2, each eligible NHANES II participant may have multiple submission records and each submission record may return one or more

⁵ 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.

potential matches to a NDI record. The NHANES II Linked Mortality File does NOT include all the potential matching NDI records. Rather for those NHANES II 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 II-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 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. Within each of the remaining potential Class 3 or 4 matches, the matches were ordered 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 NHANES II record, the tiebreaker reflected the importance of matching items⁶.

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. Next, NCHS determined whether the selected 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. For Class 3 matches, those with a score *greater than or equal* to 47 were considered true matches and for Class 4 matches, those with a score *greater than or equal* to 40 were considered true matches. Matched records below the cut-off scores were considered false matches. These scores 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. Finally, NCHS attempted to collect and review the death certificates for all deaths at or above the cut-off point to confirm mortality status.

The NHANES II – NDI Linked Mortality Data File contains 4,143 NHANES II decedents, which includes 2,124 deaths identified from the previous NHANES II mortality study. In 21 cases, the previously released death information could not be confirmed through the new NDI record match or other sources and, therefore, these cases now are assumed to be alive. Of the 4,143 deaths, 3,794 (91.5%) 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 NHANES II participant. NCHS could not verify 349 deaths with a death certificate. Among these, 222 were identified as deceased only through the NDI probabilistic match, 105 were deceased based upon death information collected in prior linkages with Medicare data, and 22 were deceased based upon the Social Security Administration's Death Master File.

⁶ 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 NHANES II participants as of 1992 Follow-up and the current Mortality Update

	NHANES II Mortality Update			
NHANES II 1992 Follow-up	Alive	Deceased	Ineligible	Total
Alive	5,085	2,019	1	7,105
Deceased	21	2,124	0	2,145
Ineligible	0	0	2	2
Total	5,106	4,143	3	9,252

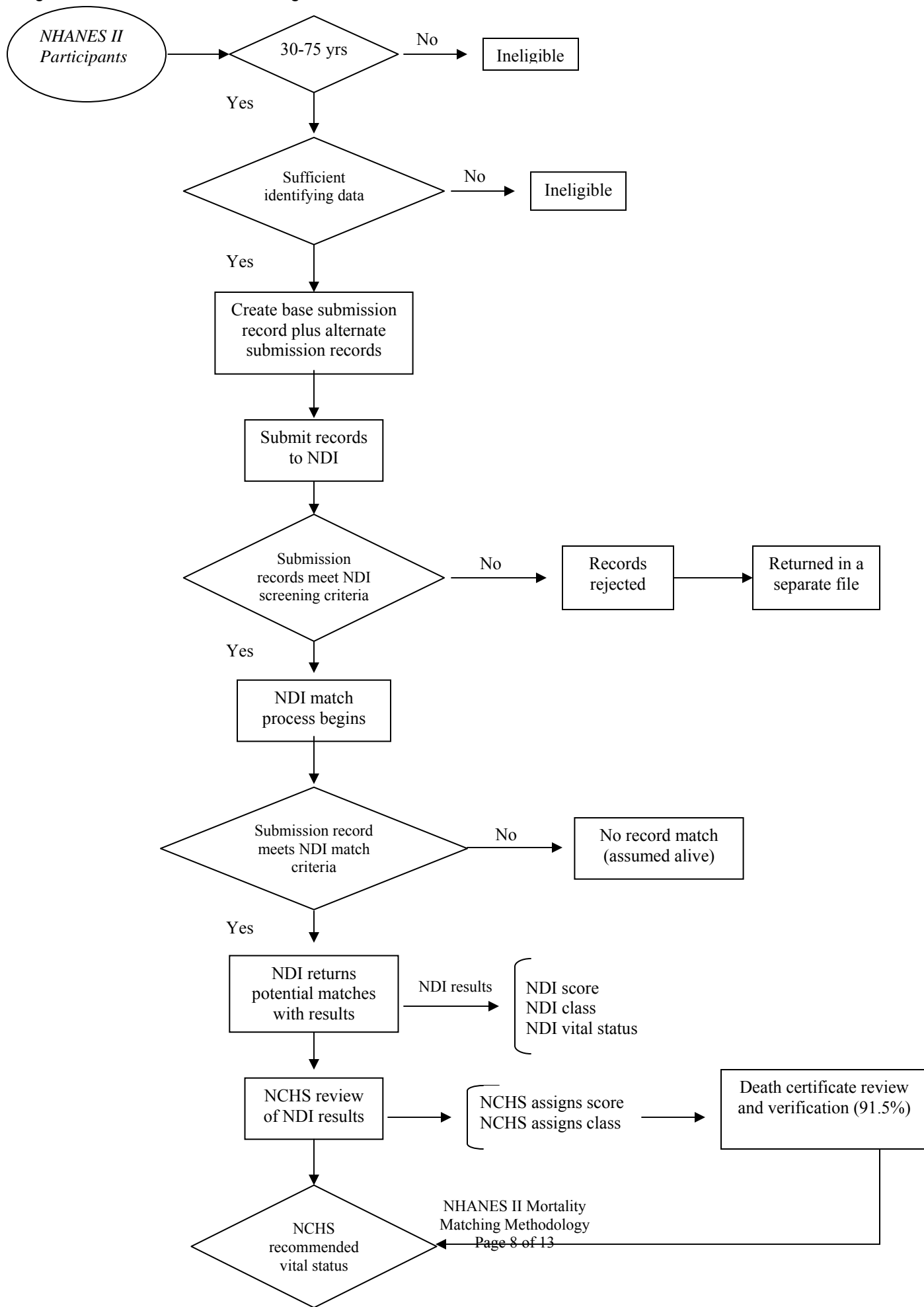
Notice to Users

Eligible NHANES II participants with a “true” NDI record match are assumed to be dead. Eligible NHANES II 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. The 3 ineligible NHANES II respondents should be excluded from mortality analyses. Analysts should use the variable MORTSTAT to determine vital status.

Users can take either a more or less conservative approach to vital status ascertainment by changing the mortality status of NHANES II participants based upon the source of mortality (MORTSRCE). The source of mortality is coded into 5 categories and is available on the standard NHANES II Linked Mortality File. For example, to be very conservative in the assignment of vital status, using the MORTSRCE variable, the user could limit the determination of deceased only to those whose death was confirmed by a death certificate. In addition, a variable indicating the vital status as of the 1992 follow-up (VITST92) has been provided on the NHANES II standard mortality file.

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 II 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 II-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 NHANES II 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 II 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 II 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 II-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

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 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 the death of 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 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 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 1971-1975. 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.

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.

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* 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 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

