The Linkage of National Center for Health Statistics Survey Data to the National Death Index – 2015 Linked Mortality File (LMF): Methodology Overview and Analytic Considerations

Data Release Date: November 6, 2017
Document Version Date: April 11, 2019

Office of Analysis and Epidemiology
National Center for Health Statistics
Centers for Disease Control and Prevention
datalinkage@cdc.gov

Suggested Citation:

National Center for Health Statistics. Office of Analysis and Epidemiology. The Linkage of National Center for Health Statistics Survey Data to the National Death Index — 2015 Linked Mortality File (LMF): Methodology Overview and Analytic Considerations, March 2019. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data-linkage/mortality-methods.htm

Contents

1 Introduction

The National Center for Health Statistics (NCHS) has developed a record linkage program designed to maximize the scientific value of the Center's population-based surveys. The linkage between the NCHS survey data and the National Death Index (NDI) is intended to maximize the scientific value of NCHS surveys without increasing respondent reporting burden. These data, collectively referred to as the Linked Mortality Files (LMFs), include mortality follow-up data through December 31, 2015.

The data used in this linkage were from the following populated-based NCHS health surveys and years:

- National Health Interview Survey (NHIS): 1985-2014
- Continuous National Health and Nutrition Examination Survey (NHANES): 1999-2014
- NHANES III (1988-1994)
- NHANES II (1976-1980)
- NHANES I Epidemiologic Follow-up Study (NHEFS)
- Second Longitudinal Study of Aging (LSOA II)
- Supplement on Aging (SOA)
- National Home and Hospice Care Survey (NHHCS): 2007
- National Nursing Home Survey (NNHS): 1985, 1995, 1997, 2004

These data were linked to the NDI, a centralized database of U.S. death records gathered from states' vital statistics offices. The NDI contains death record information for each person dying in the United States or a U.S. territory from 1979 through 2015.

This document presents an overview of the methodology and analytic considerations for researchers using the restricted-use LMFs.

For more information or questions about the LMF, please visit the data linkage website or contact the NCHS Special Projects Branch at datalinkage@cdc.gov.

2 Access to (Restricted-Use) NCHS-NDI 2015 Linked Mortality Files

The NCHS must provide safeguards for the confidentiality of its survey participants. To ensure confidentiality, all personal identifiers have been removed from the NCHS-NDI 2015 LMFs. However, there remains the small possibility of re-identification and for this reason, the NCHS-NDI LMFs are not available as public-use files. Researchers who want to obtain the NCHS-NDI LMFs must submit a research proposal to the NCHS Research Data Center (RDC) to obtain permission to access the restricted use files. All researchers must submit a research proposal to determine if their project is feasible and to gain access to these restricted data files. The proposal provides a framework which allows RDC staff to identify potential disclosure risks.

NCHS RDCs are housed on-site at Centers for Disease Control and Prevention (CDC) facilities in Hyattsville, MD, Washington, DC, and Atlanta, GA. In addition, NCHS restricted data can be accessed from RDCs housed in U.S. Census Bureau offices in several locations across the country. Researchers generally will need to be on-site at one of the RDCs to access restricted-use linked data, including the restricted-use NCHS-NDI LMFs, although remote access is permitted under certain conditions. Within the RDC, the NCHS-NDI LMFs can be merged with NCHS restricted (if needed) and public-use survey data files using unique survey person identification numbers.

3 NCHS Surveys linked to the NDI through December 31, 2015

National Health Interview Survey (NHIS)

NHIS is a nationally representative, cross-sectional household interview survey that serves as an important source of information on the health of the civilian, non-institutionalized population of the United States. It is a multistage sample survey with primary sampling units of counties or adjacent counties, secondary sampling units of clusters of houses, tertiary sampling units of households, and finally, persons within households. It has been conducted continuously since 1957 and the content of the survey is periodically updated. NHIS has been used as the sampling frame for other NCHS surveys focusing on specialized populations, including LSOA II. Prior to 2007, NHIS traditionally collected full 9-digit Social Security Numbers (SSN) from survey participants. However, in attempt to address respondents' increasing refusal to provide SSN and consent for linkage, NHIS began, in 2007, to collect only the last four digits of SSN and added an explicit question about linkage for those who refused to provide SSN. NHIS is currently planning a content and structure redesign. For detailed information on the NHIS's contents and methods, refer to the NHIS website.

National Health and Nutrition Examination Survey (NHANES)

NHANES is a continuous, nationally representative survey consisting of about 5,000 persons from 15 different counties each year. For a variety of reasons, including disclosure issues, the NHANES data are released on public-use data files in two-year increments. The survey includes a standardized physical examination, laboratory tests, and questionnaires that cover various health-related topics. NHANES includes an interview in the household followed by an examination in a mobile examination center (MEC). NHANES is a nationally representative, cross-sectional sample of the U.S. civilian, non-institutionalized population that is selected using a complex, multistage probability design.

Prior to becoming a continuous survey in 1999, NHANES was conducted periodically, with the last periodic survey, NHANES III, conducted between 1988 and 1994. NHANES III was designed to provide national estimates of health and nutritional status of the civilian, non-institutionalized population of the United States aged 2 months and older. Similar to the continuous survey, NHANES III included a standardized physical examination, laboratory tests, and questionnaires that covered various health-related topics.

NHANES II preceded NHANES III and was conducted from 1976-1980. It is also a nationwide probability sample of 27,801 persons aged 6 months to 74 years of age. As with NHANES III, the survey was designed to provide national estimates of the health and nutritional status of the civilian non-institutionalized U.S. population and included a standardized physical examination, laboratory tests, and questionnaires that covered various health-related topics. More information on NHANES II is available here. Only NHANES II survey participants that were part of the NHANES II Mortality study were eligible for inclusion in the NCHS-NDI 2015 LMF.[1]

[1] https://www.cdc.gov/nchs/data/series/sr_01/sr01_038.pdf

NHANES I Epidemiologic Follow-Up Study (NHEFS)

NHEFS was a national longitudinal study conducted in collaboration with the National Institutes of Health, National Institute on Aging and other agencies of the Public Health Service. The NHEFS cohort included all persons 25-74 years of age who completed a medical examination as part of NHANES I in 1971-75. The NHEFS study design included four follow-up interviews, conducted in 1982-84, 1986, 1987 and 1992, to investigate the relationships between clinical, nutritional, and behavioral factors assessed at baseline, and subsequent morbidity, mortality, and institutionalization.

The Second Longitudinal Study of Aging (LSOA II)

LSOA II was a prospective study of a nationally representative sample of civilian, non-institutionalized persons 70 years of age and over at the time of their 1994 NHIS interview, which served as the baseline for the study. The LSOA II study design included two follow-up telephone interviews, conducted in 1997-98 and 1999-2000. The LSOA II provides information on changes in disability and functioning, individual health risks and behaviors in the elderly, and use of medical care and services employed for assisted community living. For detailed information on the LSOA II contents and methods, refer to the LSOA II website.

Supplement on Aging (SOA)

SOA is a nationally representative sample comprised of 16,148 civilian non-institutionalized persons aged 55 years and over at the time of their 1984 National Health Interview Survey (NHIS) interview. SOA forms the basis for the prospective study, the Longitudinal Study of Aging, with the primary goal of characterizing the health and social status of people aged 55 years and over in the United States. SOA also provides information on how health factors influencing the aging individual interact with psychosocial and environmental factors; enhances the knowledge base for investigating issues of prevention and postponement of disability and dependency for the aging individual; delineates research issues related to enhancement of care, social support, and coping for older individuals who become disabled; and provides information about the factors that influence a persons' ability to live independently as they age. More information on SOA is available here.

National Home and Hospice Care Survey (NHHCS)

NHHCS is one in a series of nationally representative sample surveys of U.S. home health and hospice agencies. It is designed to provide descriptive information on home health and hospice agencies, their staffs, their services, and their patients. NHHCS was first conducted in 1992 and was repeated in 1993, 1994, 1996, 1998, and 2000, and most recently in 2007. Only the most recent year of NHHCS was included in this linkage. For more information on the NHHCS content and methods, refer to the NHHCS website.

National Nursing Home Survey (NNHS)

NNHS provides information on nursing homes from two perspectives: that of the provider of services and that of the recipient of care. Data for the surveys were obtained through personal interviews with facility administrators and designated staff who used administrative records to answer questions about the facilities, staff, services and programs, and medical records to answer questions about the residents. NNHS was first conducted in 1973-1974 and repeated in 1977, 1985, 1995, 1997, 1999, and most recently in 2004. The 1985, 1995, 1997, and 2004 surveys were included in the NCHS-NDI linkage. For more information on the NNHS content and methods, refer to the NNHS website.

4 Matching Methodology Overview

Mortality status for NCHS survey participants was ascertained primarily through probabilistic record matching with the NDI. The NCHS Special Projects Branch (SPB) employed a matching methodology for the 2015 LMF 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. These sources include information obtained from linkages with the Social Security Administration, the Centers for Medicare and Medicaid Services, as well as death certificate or data collection information obtained from previous follow-up studies (e.g., NHEFS, NHANES II, LSOA II, and 1985 NNHS). Participant vital status took into account both the NDI matching record and mortality status derived from other source(s). See Section 5.6 for additional details.

4.1 Changes from Previous Releases

The previous NCHS-NDI LMF provided mortality data through December 31, 2011. The updated 2015 LMF supersedes the previous linkages of NCHS survey data linked to the NDI.

This 2015 version of the LMF 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 four additional years (through December 31, 2015) of mortality follow-up; (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; and (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 years of the surveys.

4.2 Overview of Linkage Process

The linkage of NCHS survey participant's data to the NDI involves six broad steps. Briefly, they are:

1. Identify participants from the NCHS surveys and determine participant eligibility.
2. Create base submission record plus any alternate records.[2]
3. Merge base submission file with NDI data to create record selection file.
4. Execute match process.
5. Review match results.
6. Select matches and determine vital status.

[2] 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 a 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.

Details of these steps are beyond the scope of this document but they will be provided in a forthcoming NCHS Series 1 Report.

4.3 Linkage Fields

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

Social Security Number (SSN)
First Name
Middle Initial
Last Name
Father's Surname
Month of Birth
Day of Birth
Year of Birth
State of Birth
State of Residence
Sex
Race
Marital Status

Collection of the data necessary to conduct the NCHS-NDI linkage differed by NCHS surveys:

- The NHIS has routinely collected all of the data items used for the linkage. However, beginning in 2007 only the last four digits of the SSN — not the full nine digits — were collected.
- NHEFS, NHANES III, continuous NHANES, LSOA II, and NHHCS 2007 collected all of the data items.
- NHANES II did not collect SSN but SSN was obtained through secondary data collection as noted in Section 3.
- NNHS and SOA collected most of the data items used for the linkage: 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 alternative records as noted in Section 4.2. NDI records could then be matched to any or all of the submission records created for a survey participant.

All participants with sufficient identifying data were eligible for mortality linkage. 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), last name, first name
2. SSN (nine digits or last four digits), sex, month of birth, day of birth, year of birth
3. Last name, first name, month of birth, year of birth

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 NNHS public-use files but these participants were ineligible for the NDI linkage (for reasons noted above) and are thus indicated as ineligible on the 2015 LMFs.

4.4 Identifying Potential Match Records

Potential NDI death record matches are initially based on various combinations of matching identifiers between the files. Similar to previous linkages, NCHS SPB applied the criteria used by the NDI to identify potential matches (see Appendix A of the NDI User's Manual). Briefly, the criteria include various combinations of Social Security Number; first name/initial, last name/initial, and birth surname; exact and approximate year of birth; exact month of birth; exact day of birth; sex; race; and state of birth.

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 NCHS survey submission record that matched a NDI record on the specified criteria was selected as a potential match. The NDI record selection process could return several potential matches for each person resulting in some non-matches or duplicate records.

4.5 Scoring and Classifying Potential Match Records

There are several ways that a NCHS survey participant submission record could match to a NDI record. For each potential match, a flag was created to indicate whether there was agreement or disagreement 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 frequency 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 prevalence 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.

Scoring differs depending on whether four or nine digit SSNs are available.

$$\begin{aligned}\text{Score}_{\text{9 digit SSN}} = {} & \{W_{\text{SSN 1}} + \ldots + W_{\text{SSN 9}}\} + (p)W_{\text{first name X sex X birth year}} \\ & + W_{\text{middle initial X sex}} + (q)W_{\text{last name}} + W_{\text{race}} + W_{\text{sex}} \\ & + W_{\text{marital status X sex X age}} + W_{\text{birth day}} + W_{\text{birth month}} + (r)W_{\text{birth year}}\end{aligned}$$

$$\begin{aligned} &+ W_{\text{state of birth}} + W_{\text{state of residence}} + W_{\text{survey middle / NDI first}} \\ &+ W_{\text{survey middle initial / NDI first initial}} + W_{\text{surname (female)}} \end{aligned}$$

where

W = weight,
p, q, r = proration factors for inexact matches (including matches by NYSIIS/Soundex, initials, and year of birth (± 3 years)), and
SSN_i = i^{th} digit of the SSN.

For surveys that collected nine digits of SSN, a record needed to agree on at least eight digits of SSN to be assigned the maximum weight for SSN. 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 the years of NHIS 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 SSNs a slightly higher score for matching on the last four digits without giving an equivalent score to records with nine digit SSNs.

$$\begin{aligned} \text{Score}_{\text{4 digit SSN}} = \{ &W_{SSNa} + W_{SSN\ 6} + \ldots + W_{SSN\ 9}\} + (p)W_{\text{first name X sex X birth year}} \\ &+ W_{\text{middle initial X sex}} + (q)W_{\text{last name}} + W_{\text{race}} + W_{\text{sex}} \\ &+ W_{\text{marital status X sex X age}} + W_{\text{birth day}} + W_{\text{birth month}} + (r)W_{\text{birth year}} \\ &+ W_{\text{state of birth}} + W_{\text{state of residence}} + W_{\text{survey middle / NDI first}} \\ &+ W_{\text{survey middle initial / NDI first initial}} + W_{\text{surname (female)}} \end{aligned}$$

where W, p, q, r, and SSN_i are as above.

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 decisive for determining true matches than others (e.g., SSN versus state of birth) and that non-changing identifying information is more substantial 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 noted below.

As SSN is a key identifier in the matching process, each NCHS-NDI record match was initially classified according to whether a four or nine 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 2015 Linked Mortality file were as follows.

Class 1: Agreed on at least eight (of nine) or four (of four) digits of SSN, first name (including NYSIIS/Soundex agreement), middle initial (including blank), last name (including NYSIIS/Soundex agreement), birth year (±3 years), birth month, sex, and state of birth.

Class 2: Agreed on at least seven (of nine) or four (of four) digits of SSN and at least five of the following items: first name (including NYSIIS/Soundex agreement), middle initial (including blank), last name (including NYSIIS/Soundex agreement), 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 seven of the following items agreed: first name (including NYSIIS/Soundex match), middle initial (including blank), 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 either SSN was recorded incorrectly or that the spouse's SSN was recorded instead of the subject's SSN. In this category, SSN was known but three or more (of nine) and one or more (of four) digits did not agree, but at least eight of the following items agreed: first name (including NYSIIS/Soundex match), middle initial (including blank), last name (including NYSIIS/Soundex match), birth year (±3 years), 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 eight of the items listed in either of the Class 3 types matched.

Class 5: SSN was present but fewer than 7 (of 9) or 4 (of 4) digits on SSN agreed, and the record did not fall into a prior class.

4.6 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 a 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 less than or equal to zero were considered false matches and were eliminated from the pool of potential matches. 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 (Note: Class 5 records were determined non-matches). 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 matching criteria were compared and the record with the most matching criteria was selected.

Next, NCHS SPB determined whether each best record was a match. A match reflects both the best match for vital status of the survey participant and a corresponding match to the correct death certificate data. All class 1 match records were considered matches. Within each class, matches with a score greater than or equal to the cut-off score were considered matches whereas records with a score less than the cut-off were considered non-matches. The cut-off scores for classes 2, 3, and 4 were 44, 45, and 42, respectively. The cut-off scores within classes 2, 3, and 4 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.

5 Analytic Considerations

5.1 Linkage Eligibility Status

All participants with sufficient identifying data were eligible for mortality follow-up. Each record was screened to determine if it contained at least one of the combinations of identifying data elements listed in Section 4.3.

Any survey participant record that did not meet the minimum data requirements was ineligible for record linkage.

Eligibility status for mortality follow-up is indicated by the variable ELIGSTAT. For analyses using the linked mortality files, analysts should limit their analysis to those survey records with a value of ELIGSTAT = 1. On average, 94.8% of the survey participants were eligible for the mortality follow-up.

5.2 Use of Survey Weights

Survey Sampling Weights

The use of sampling weights and sample design variables is recommended to account for the complex survey design of NCHS studies. Failure to account for the complex survey design may produce biased estimates and overstated significance levels.

Eligibility Adjusted Sampling Weights

For analyses using the LMF, researchers should consider adjusting the original sampling weight to account for those ineligible for linkage to the NDI due to insufficient identifying data. Ignoring those ineligible for linkage in the 2015 LMFs may lead to biased mortality estimates. NCHS has provided guidance on methods available to adjust sampling weights: Use of Survey Weights for Linked Data Files – Preliminary Guidance.

NHIS Eligibility Adjusted Sample Weights

The NCHS Special Projects Branch has provided eligibility adjusted weights for the NHIS from 1987-2014. For the 2015 Linked Mortality Files, there are no eligibility adjusted sample weights for the 1985 and 1986 NHIS. NCHS recommends using the public-use annual final basic weight (WTFA) for those survey years.

For the 1987-2014 NHIS, participants that are classified as eligible for mortality follow-up had their original NHIS sampling weight adjusted to account for those ineligible for linkage to the NDI due to insufficient identifying data provided in the survey. The new eligibility adjusted sample weights provided on the 2015 LMF are recommended for use in place of the original NHIS sample weights to prevent biased mortality estimates.

The 2015 LMFs include three eligibility adjusted sample weights for the NHIS:

- WGT NEW refers to a person-level record and is available for the NHIS years 1987-2014
- SA WGT NEW refers to a sample adult record and is available for the NHIS years 1997-2014
- SC WGT NEW refers to a sample child record and is available for the NHIS years 1997-2014

The NHIS from 1987-1996 did not include sample adult or sample child files.

Technical note: Treating the eligible sample from the NHIS as a subsample of the original NHIS sample allows for the original post-stratification adjustment method to be used to inflate the sampling weights. The tacit assumption is the adjustment cells used will mitigate estimation bias due to using only the eligible sample.

5.3 Pooled Analysis of NCHS Linked Mortality Files

To increase the sample size for many types of analyses, analysts may wish to pool several survey years (or cycles). When survey years (cycles) are combined, the estimates will be representative of the population at the midpoint of the combined survey period. Analysts should refer to the specific surveys (e.g., NHIS, NHANES) regarding how to adjust sample weights when pooling years. A simple, valid weight adjustment procedure that NCHS often recommends is to divide each sample weight in the pooled dataset by the number of years that are being pooled. For example, divide by 2 when two years (cycles) of survey data are combined, divide by 3 when three years of data are combined, etc.

Please note that when combining survey years (cycles) it is the data users' responsibility to examine possible changes in variable names and/or locations of the data files. Differences in study design variables may also be an issue when pooling survey years within a specific survey.

Pooled Analysis Method for Estimating Variance

NHIS has provided analysts with guidance for variance estimation for pooled analyses of NHIS years. Please refer to the following NHIS file documentation for additional information:

- NHIS 1986-1994
- NHIS 1995-1996
- NHIS 1997-2005
- NHIS 2006-2015

NHANES also provides tutorials on pooling years of NHANES data, including construction of appropriate pooled sample weights. Links to the NHANES tutorials can be found here.

5.4 Linkage of Survey Participants with Improbable Ages

The NCHS 2015 LMFs include records where the calculated age presumed alive at the end of mortality follow-up is 100 years or more. For these cases, there was no valid NDI record match or any other source of mortality information. Given the probabilistic nature of the mortality ascertainment and the lower likelihood of being alive at 100 years or older, analysts may wish to consider these cases as loss to follow-up and make them ineligible for mortality analyses. (Note: NDI only includes deaths that occurred in the United States or a U.S. territory and therefore may not include deaths of all survey participants.)

A practical method for determining an age cutoff at which participants should be considered lost to follow-up is to use the probability of a member in a particular population dying at, or living to, a particular age. The Social Security Administration (SSA) published a report in 2005 containing projections of mortality for cohorts of births in decennial years 1900 through 2100.[3] Based on these cohort life tables, the NCHS Special Projects Branch calculated probabilities of death, conditional on year of birth and sex, but not adjusted for last known alive year (typically the year of survey response). These probabilities are available for researchers upon request. Please refer to the SSA report for more information.

5.5 Inconsistencies in Baseline Age and Follow-Up Age

Misreporting or discrepancies between reported age at interview and the date of birth may result in values for age at death or age last presumed alive that are inconsistent with baseline age, resulting in negative follow-up time for survival analyses. The number of cases where this occurs is small but analysts should be aware and make appropriate adjustments to the data. Researchers are encouraged to look at the source of death since, in most of the instances, date of death information was obtained through a non-NDI source (see below).

5.6 Source of Mortality Information

The primary determination of mortality for eligible participants is based upon matching survey records to the NDI, although additional sources are also incorporated. These sources include the Social Security Administration, the Centers for Medicare and Medicaid Services, data collection, and for NCHS' follow- up surveys (e.g., NHEFS), ascertainment of death certificates. If a source of mortality other than NDI was available the participant was considered deceased. Variables indicating which source, or sources, were used to determine vital status are included in the 2015 LMF Data Dictionary.

5.7 Restricted-Use Linked Mortality Files Match Result Variables

If analysts want to alter the criteria for determining final vital status, NCHS provides the match score and match class — calculated values used to determine vital status — variables to researchers on the restricted-use LMF. (Please refer to Section 4.5 for details about class and score.) Inclusion of these variables allows the analyst to conduct sensitivity analyses of vital status, i.e., shifting the cut-off score within class to be more or less conservative with respect to declaring vital status.[4]

In addition, variables that indicate the components of the matching algorithm that matched to the NDI are also available to researchers using the restricted-use LMF. These variables are noted as the NDI record match variables (prefixed with NDI_*), and include NDI record match results for all survey participants who returned a potential NDI match record, independent of whether

[3] Life Tables for the United States Social Security Area 1900-2100. SSA Pub No. 11-11536. Available at https://www.ssa.gov/oact/NOTES/pdf_studies/study120.pdf.

[4] Researchers interested in pursuing this type of analysis may find the following reference useful: Lariscy JT. Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox. *J Aging Health*. 2011 23(8):1263-84.

the participant's vital status was determined dead by the matching algorithm. For example, if the first name matched then the value in the variable NDI_FIRST would be X. These variables allow researchers to assess NDI match results for survey participants whose final vital status was determined to be alive and to conduct sensitivity analyses among decedents. The complete list of NDI_* variables is contained in the data dictionary, Death Certificate and NDI Match Variables, accessible from the NCHS Data Linkage Restricted-Use LMF webpage.

5.8 Statistical Issues with Linked Mortality Data

Analysis of the LMF presents unique statistical issues due to the potential for differential follow-up times and censoring bias. For an overview of these issues, please refer to the following report: Statistical Issues in Analyzing the NHANES I Epidemiologic Follow-up Study.

5.9 1992 NHIS Hispanic Oversample

There are NHIS participants in the 1991 NHIS sample who are also included the 1992 NHIS sample. If a researcher plans to pool these two years of survey data for their analysis, it is recommended to use the special 1992 NHIS file that excludes the participants who were also interviewed in 1991. For more information, please refer to the NHIS public-use data documentation supplement: 1992 Core Files – Version without Hispanic Oversample. Analysts are advised to take the Hispanic oversample under consideration if they combine survey years with the 1992 NHIS. In addition, if researchers exclude the participants who were also interviewed in 1991 they will need to create new adjusted weights for ineligible respondents (WGT_NEW). Guidance for the construction of new weights can be found in Appendix III of this series report.

5.10 1985 NHIS

The public-use 1985 NHIS sample includes a subset of survey participants that are not included on the restricted-use LMF. These participants were not eligible for linkage to the NDI. When merging the public-use NHIS data with the restricted-use LMF, the subset not on the LMF should be re-coded as ineligible for linkage (e.g., eligstat=3).

5.11 Merging Restricted-Use LMF Data and Public-Use NCHS Survey Data

The data provided on the NCHS 2015 LMFs can be merged with the NCHS public-use survey data files using unique identification (ID) numbers. However, the unique ID is different across surveys and years. Guidance on constructing the appropriate identifier (PUBLICID) in the public-use datasets for NHIS, SOA, and LSOA II is presented by survey year. (SOA and LSOA II are both supplements of the NHIS and therefore, construction of PUBLICID applies to these surveys as well.)

For each of the survey groups presented below, the locations, lengths, and descriptions of the variables in the NHIS public-use datasets required for creating PUBLICID are listed. When combining across NHIS years it is necessary to include the NHIS year as part of PUBLICID so each survey grouping includes some version of year (e.g. YEAR, SRVY_YR).

Note that the NHIS SAS input statements available from the NHIS public-use data website do **not** input all of the variables as character so converting them to character first may be necessary. Example SAS and Stata code to create PUBLICID are also provided.

The PUBLICID variable on the restricted-use LMF is in character format with an assigned length of 14.

1985-1994 NHIS

The data items QUARTER x PSU x WEEK x SEGMENT x HOUSEHOLD NUMBER x PERSON NUMBER identify a person within each NHIS year.

Table 5.1: 1985-1994 NHIS

Variable	NHIS Public-Use File Location	Variable Length	Description
YEAR	3-4	2	Year of Interview
QUARTER	5	1	Calendar Quarter of Interview
PSUNUMR	6-8	3	Random Recode of PSU
WEEKCEN	9-10	2	Week of Interview within Quarter
SEGNUM	11-12	2	Segment Number
HHNUM	13-14	2	Household Number
PNUM	15-16	2	Person Number within Household

SAS:

```
length PUBLICID $14;
PUBLICID = trim(left(YEAR||QUARTER||PSUNUMR||WEEKCEN||SEGNUM||HHNUM||PNUM));
```

Stata:

```
egen PUBLICID = concat(YEAR QUARTER PSUNUMR WEEKCEN SEGNUM HHNUM PNUM)
```

1995-1996 NHIS

The data items HOUSEHOLD NUMBER x PERSON NUMBER identify a person within each NHIS year.

Table 5.2: 1995-1996 NHIS

Variable	NHIS Public-Use File Location	Variable Length	Description
YEAR	3-4	2	Year of Interview
HHID	5-14	10	Household Number
PNUM	15-16	2	Person Number within Household

SAS:

```
length PUBLICID $14;
PUBLICID = trim(left(YEAR||HHID||PNUM));
```

Stata:

```
egen PUBLICID = concat(YEAR HHID PNUM)
```

1997-2003 NHIS

The data items HOUSEHOLD NUMBER x FAMILY NUMBER x PERSON NUMBER identify a person within each NHIS year.

Table 5.3: 1997-2003 NHIS

Variable	NHIS Public-Use File Location	Variable Length	Description
SRVY_YR	3-6	4	Year of Interview
HHX	7-12	6	Household Number
FMX	13-14	2	Family Number
PX	15-16	2	Person Number within Household

SAS:

```
length PUBLICID $14;
PUBLICID = trim(left(SRVY_YR||HHX||FMX||PX));
```

Stata:

```
egen PUBLICID = concat(SRVY_YR HHX FMX PX)
```

2004 NHIS

The data items HOUSEHOLD NUMBER x FAMILY NUMBER x PERSON NUMBER identify a person within each NHIS year.

Table 5.4: 2004 NHIS

Variable	NHIS Public-Use File Location	Variable Length	Description
SRVY_YR	3-6	4	Year of Interview
HHX	7-12	6	Household Number
FMX	13-14	2	Family Number
FPX	15-16	2	Person Number within Household

SAS:

```
length PUBLICID $14;
PUBLICID = trim(left(SRVY_YR||HHX||FMX||FPX));
```

Stata:

```
egen PUBLICID = concat(SRVY_YR HHX FMX FPX)
```

2005-2014 NHIS

The data items HOUSEHOLD NUMBER x FAMILY NUMBER x PERSON NUMBER identify a person within each NHIS year.

Table 5.5: 2005-2014 NHIS

Variable	NHIS Public-Use File Location	Variable Length	Description
SRVY_YR	3-6	4	Year of Interview
HHX	7-12	6	Household Number
FMX	16-17	2	Family Number
FPX	18-19	2	Person Number within Household

SAS:

```
length PUBLICID $14;
PUBLICID = trim(left(SRVY_YR||HHX||FMX||FPX));
```

Stata:

```
egen PUBLICID = concat(SRVY_YR HHX FMX FPX)
```

SOA / LSOA II

The data items QUARTER x PSU x WEEK x SEGMENT x HOUSEHOLD NUMBER x PERSON NUMBER identify a SOA / LSOA II participant.

Table 5.6: SOA / LSOA II

Variable	NHIS Public-Use File Location	Variable Length	Description
YEAR	3-4	2	Year of Interview
QUARTER	5	1	Calendar Quarter of Interview
PSU	6-8	3	Random Recode of PSU
WEEKPROC	9-10	2	Week of Interview within Quarter
SEGNUM	11-12	2	Segment Number
HHNUM	13-14	2	Household Number
PNUM	15-16	2	Person Number within Household

SAS:

```
length PUBLICID $14;
PUBLICID = trim(left(YEAR||QUARTER||PSU||WEEKPROC||SEGNUM||HHNUM||PNUM));
```

Stata:

```
egen PUBLICID = concat(YEAR QUARTER PSU WEEKPROC SEGNUM HHNUM PNUM)
```

6 Death Certificate Information

Additional data, obtained from the death certificate, are available to researchers using the restricted-use LMF. (The National Vital Statistics System (NVSS), formerly the Division of Vital Statistics (DVS), provides these data.) These variables are prefixed with DVS_* and are populated for different years of death year. Figures 6.1–6.6 show the availability of each of the DVS_* variables by death year across five broad variable groups (General, Occurrence Location, Residence, Other Medical Items, Decedent Characteristics). An X indicates that the variable is available for that death year; a blank indicates that it is not available. The data dictionary, Death Certificate and NDI Match Variables, on the 2015 Restricted-Use Linked Mortality File webpage contains the complete list of variable names, labels, and other metadata as described in Section 5.7.

If more information (e.g., definition of values) is sought about these variables, please refer to the NVSS Public Use Data File Documentation webpage.

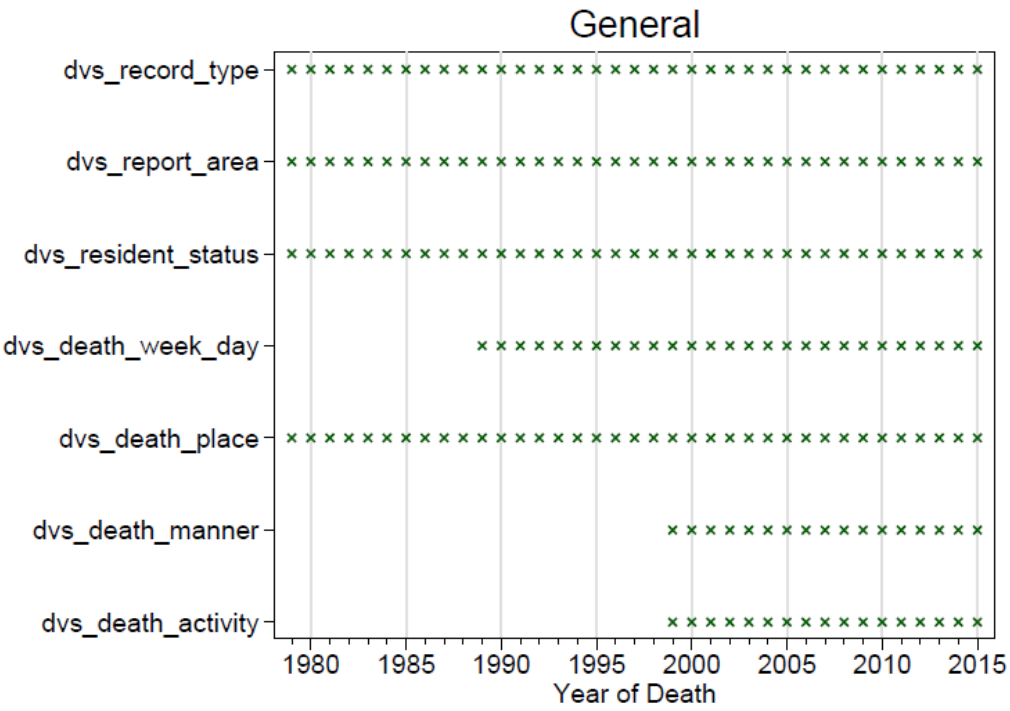

Figure 6.1: General Items: Variable Availability by Death Year

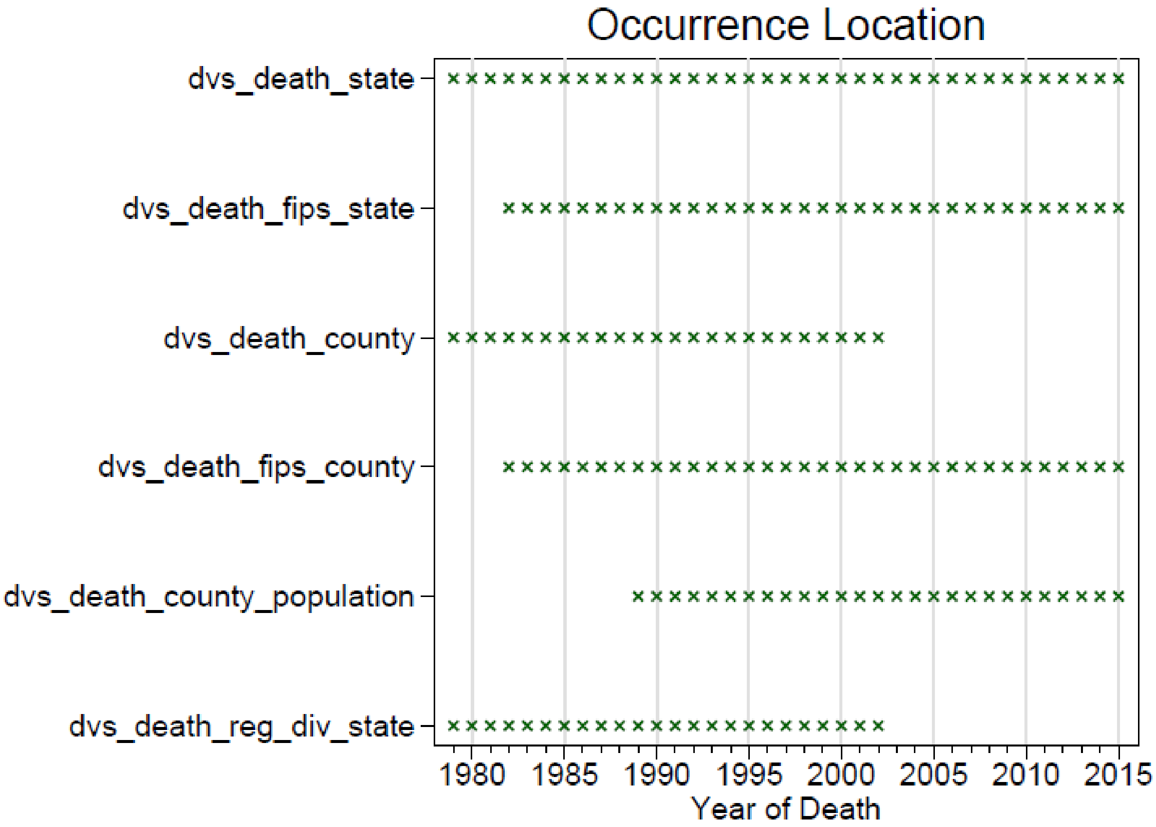

Figure 6.2: Occurrence Location: Variable Availability by Death Year

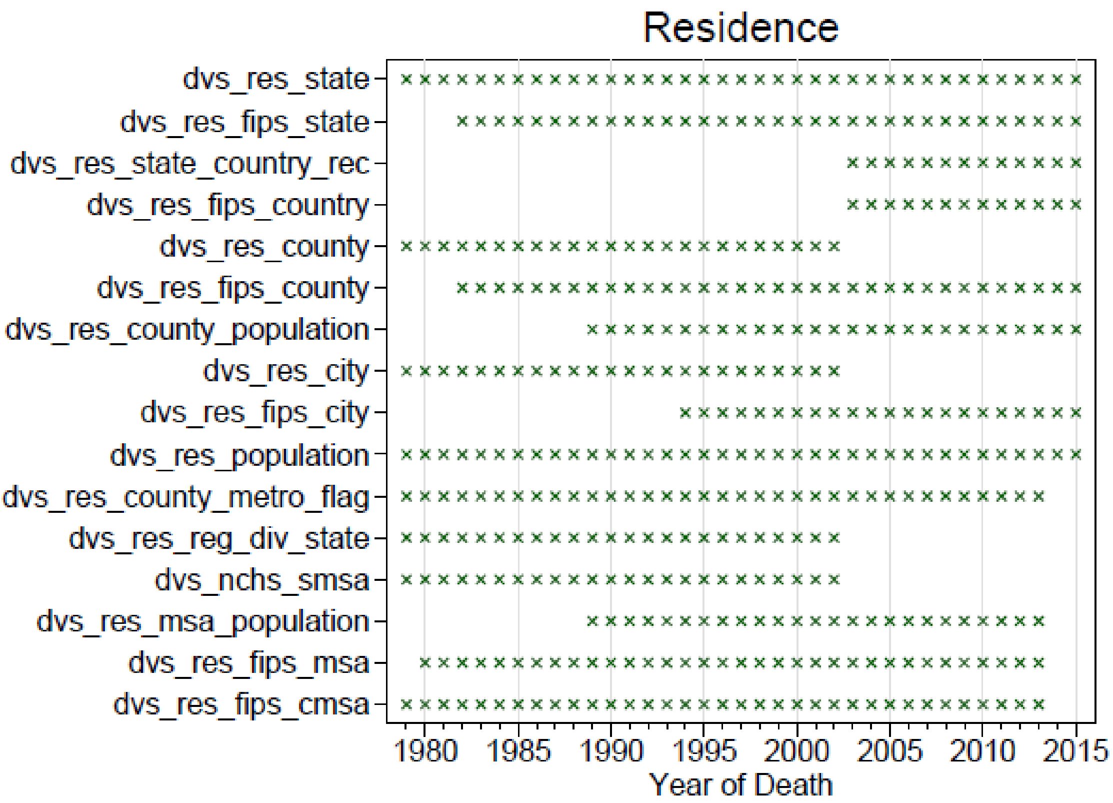

Figure 6.3: Residence: Variable Availability by Death Year

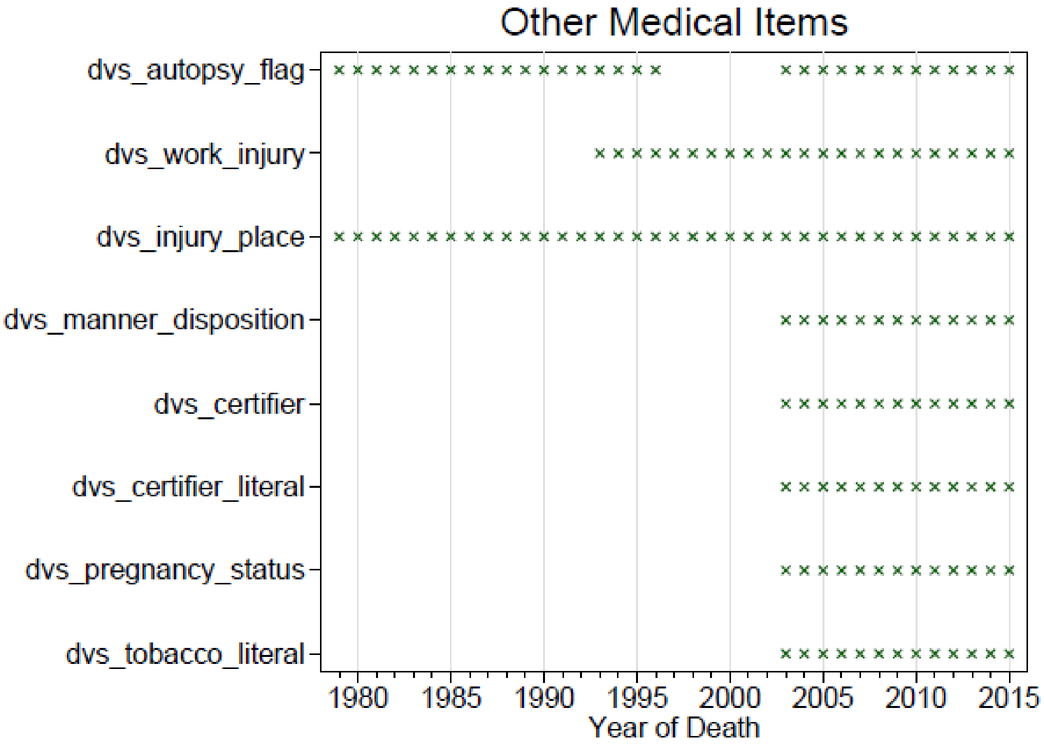

Figure 6.4: Other Medical Items: Variable Availability by Death Year

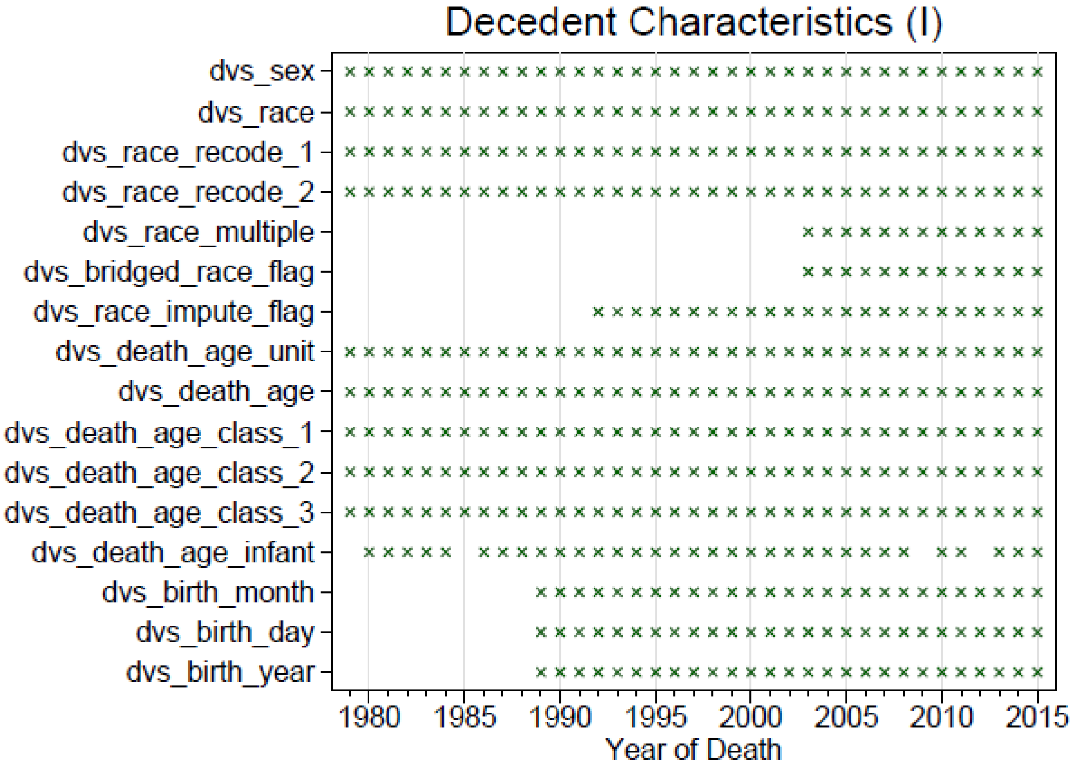

Figure 6.5: Decedent Characteristics (I): Variable Availability by Death Year

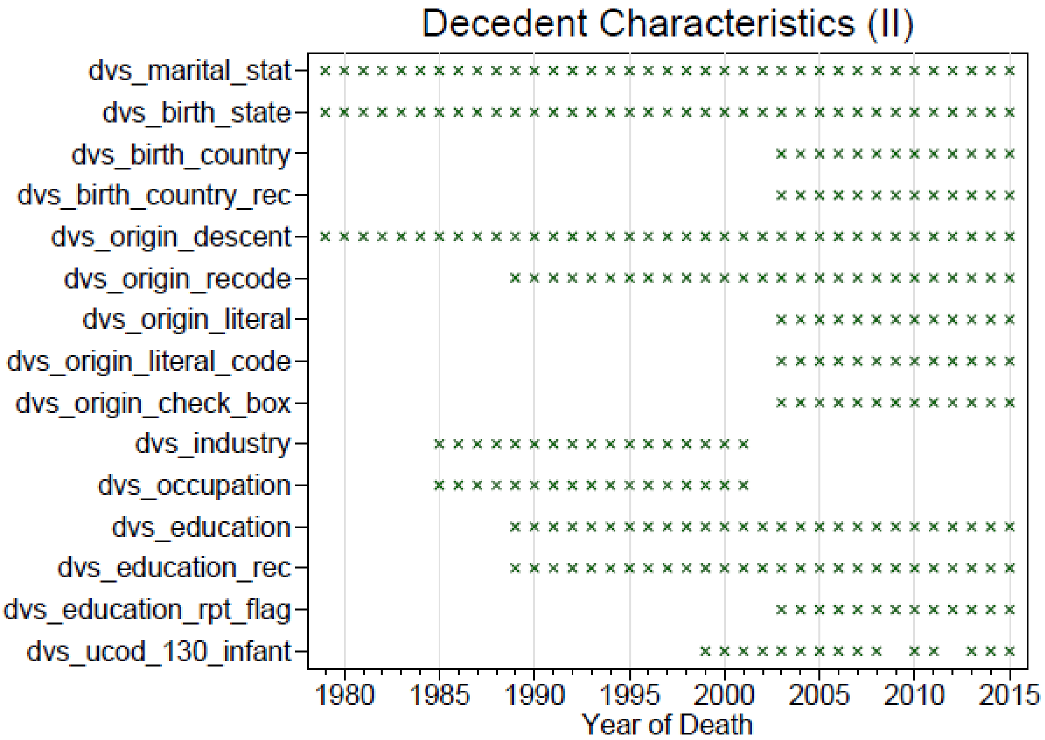

Figure 6.6: Decedent Characteristics (II): Variable Availability by Death Year