The Linkage of National Center for Health Statistics Survey Data to the National Death Index - 2022 Linked Mortality File: Linkage Methodology and Analytic Considerations

Data Release Date: January 14, 2026

Document Version Date: January 14, 2026

Division 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. Division of Analysis and Epidemiology. The Linkage of National Center for Health Statistics Survey Data to the National Death Index - 2022 Linked Mortality File: Linkage Methodology and Analytic Considerations, January 2026. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/linked-data/mortality-files/index.html

Contents

List of Acronyms

Abbreviation	Definition
CMS	Centers for Medicare & Medicaid Services
EM	expectation maximization
ERB	Ethics Review Board
HUD	U.S. Department of Housing and Urban Development
LMF	linked mortality file
NCHS	National Center for Health Statistics
NDI	National Death Index
NHANES	National Health and Nutrition Examination Survey
NHANES III	Third National Health and Nutrition Examination Survey
NHIS	National Health Interview Survey
NVSS	National Vital Statistics System
PII	personally identifiable information
RDC	Research Data Center
SSA	Social Security Administration
SSN	Social Security Number
SSN9	nine-digit Social Security Number
SSN4	four-digit Social Security Number
UCOD	underlying cause of death
VA	Department of Veterans Affairs
WTFA	public-use annual final basic weight

1 Introduction

As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to provide statistical information that can be used to guide actions and policy to improve the health of the American people. As part of its ongoing efforts to fulfill this mission, NCHS conducts several population-based and establishment surveys. Although survey data provide information on a wide-range of health-related topics, they often lack information on longitudinal outcomes.

Through its data linkage program, NCHS has been able to enhance the survey data it collects by supplementing survey information with information from mortality data from death certificates from the National Death Index (NDI). These data, collectively referred to as the Linked Mortality Files (LMFs), include mortality follow-up data through December 31, 2022.

This report describes the most recent linkage conducted between selected NCHS surveys and mortality information. A brief overview of the data sources, the methods used for linkage, and analytic guidance are provided in this report. Detailed information on the linkage methodology is provided in Appendix I. For more information or questions about the LMFs, please visit the data linkage website or contact the NCHS Data Linkage Program at datalinkage@cdc.gov.

2 Data Sources

This section provides a brief description of the NCHS surveys included in the linkage and the NDI.

2.1 National Center for Health Statistics Survey Data

The data used in this linkage were from the following NCHS population-based surveys:

- 1986-2021 National Health Interview Survey (NHIS)
- 1999-2018 Continuous National Health and Nutrition Examination Survey (NHANES)
- 2017-March 2020 Pre-Pandemic NHANES
- Third National Health and Nutrition Examination Survey (NHANES III)

2.1.1 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, noninstitutionalized 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.

Prior to 2007, NHIS collected full 9-digit Social Security Numbers (SSN) from survey participants. In 2007, to address increasing refusal to provide SSN and consent for linkage, NHIS began to collect only the last 4 digits of SSN and added an explicit question about linkage for those who refused to provide SSN. Also, in 2019 NHIS implemented its most recent content and structure redesign. For detailed information on the NHIS's contents and methods, refer to the NHIS website, https://www.cdc.gov/nchs/nhis/index.html.

2.1.2 National Health and Nutrition Examination Survey (NHANES)

NHANES is a continuous, nationally representative cross-sectional survey designed to monitor the health and nutritional status of the civilian noninstitutionalized U.S. population. The NHANES sample is selected through a complex, multistage probability design. The sample design includes oversampling to obtain reliable estimates of health and nutritional estimates for population subgroups. The survey consists of interviews conducted in participants' homes and standardized health examinations conducted in mobile examination centers.

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

Due to the coronavirus disease 2019 (COVID-19) pandemic the NHANES program suspended field operations in March 2020. Data collected from 2019 to March 2020 were combined with data from the NHANES 2017-2018 cycle to form a nationally representative sample of NHANES 2017-March 2020 pre-pandemic data. For more information about the NHANES 2017-March 2020 pre-pandemic file, refer to the NHANES website https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020.

NHANES continued to collect full nine-digit SSN through the 2017-2018 survey cycle. Starting in 2017-2018, survey participants who consented to linkage but who refused to provide their full nine-digit SSN were given the option to provide only the last four digits.

For detailed information about the Continuous NHANES and NHANES III contents and methods, refer to the NHANES website, https://www.cdc.gov/nchs/nhanes/.

2.2 National Death Index (NDI)

The NDI is a centralized database of United States death record information that NCHS established working with states. The NDI is a resource to help the public health and medical research community with their mortality ascertainment activities. The NDI became operational in 1981 and includes death record information for persons who died in the U.S. or a U.S. territory from 1979 onward. The records, which are compiled annually, include detailed information on the underlying and multiple causes of death. **This linkage with the NDI contains deaths from January 1, 1986 through December 31, 2022.**

The NDI contains identifying information for each death that can be used to conduct linkages. The identifiers from the NDI that are used in the linkage with the survey data are SSN, first name, middle initial, last name, father's surname, month of birth, day of birth, year of birth, sex, and state of residence. For more information about the NDI, refer to the NDI website, https://www.cdc.gov/nchs/ndi/index.html.

3 Linkage Methodology

This section outlines steps that were used to link NCHS survey data to NDI data.

3.1 Linkage Eligibility Determination

The linkage of the NCHS survey data and the NDI was reviewed and approved by the NCHS Research Ethics Review Board (ERB)[1]. All survey participants with sufficient identifying data were eligible for mortality linkage. Each survey participant's record was screened to determine if it contained at least one of the following combinations of identifying data elements:

- SSN[2] (nine digits (SSN9), or last four digits (SSN4)), last name, first name
- SSN[2] (nine digits (SSN9), or last four digits (SSN4)), sex, month of birth, day of birth, year of birth
- Last name, first name, month of birth, year of birth

Any survey participant records that did not meet these minimum data requirements were considered ineligible for record linkage. **Note**: For NHIS, beginning in 2015, detailed information required for linkage was collected only for the sample adult and sample child survey participants. While the NHIS collects some information about all members of the household, only sample adult and sample child participants were eligible for mortality linkage.

The variable ELIGSTAT, included on the linked survey-NDI mortality variables file, provides the linkage eligibility status for each survey record: ELIGSTAT values include 0 (ineligible) or 1 (eligible).

3.2 Linkage Overview

Person identifiers (e.g., SSN, name, date of birth) collected by the surveys were used to link NDI mortality records for survey participants who died through the end of the 2022 calendar year. For more detailed information on linkage methodology see Appendix I: Detailed Description of Linkage Methodology.

Linkage-eligible survey records were linked to the NDI using the following identifiers: SSN, first name, last name, middle initial, month of birth, day of birth, year of birth, state of residence, and sex.

The NCHS survey participant records and the NDI were linked using both deterministic and probabilistic approaches and were conducted separately for males and females[3].

[1] The NCHS ERB is an appointed ethics review committee that is established to protect the rights and welfare of human research subjects.

[2] Nine-digit SSN is considered valid if: 9-digits in length, containing only numbers, does not begin with 000, 666, or any values after 899, all 9-digits cannot be the same (i.e., 111111111, etc.), middle two and last 4-digits cannot be 0's (i.e., xxx-00-xxxx or xxx-xx-0000), and digits are not consecutive (ex. 012345678). Additionally, special SSN values (i.e., 123-123-1234, 111-22-3333, 010-010-0101, 001-01-0001, etc.) were changed to missing. Four-digit SSN is considered valid if: 4-digits in length, containing only numbers, and is between 0001 and 9999.

[3] Because first names are commonly associated with a person's sex, conducting the linkage separately for males and females helps to ensure independence and more appropriate weighting of name comparisons. Additionally, multiple part first and last names are more likely to be associated with females, which are handled differently when creating the linkage submission file. See Appendix I, Section 1 for additional information on the alternate record generation process for multiple part names.

1. Deterministic linkage joined records on exact SSN (nine digits or last four digits) and validated links by comparing other identifying fields (i.e., first name, last name, day of birth, etc.)

2. Probabilistic linkage identified likely matches, or links, between all records. All records were probabilistically linked[4] and scored as follows:
    a. Formed pairs via blocking
    b. Scored pairs
    c. Modeled probability - assigned estimated probability that pairs are links

3. Pairs were selected that were believed to represent the same individual between data sources (i.e., they are a match).
    a. Deterministic matches (from step 1) were assigned a match probability of 1
    b. Record pairs selected from the probabilistic match (step 2) were assigned the model match probability. Record pairs with a match probability above the probability cut-off value were determined to be matches.

3.3 Linkage Rates

Tables 1 and 2 provide linkage eligibility rates and linkage results for adults 18 and over. For each of these linked NCHS surveys, the tables present the total survey sample size, the sample size eligible for the NDI linkage, the number of eligible survey participants linked to the NDI, and the match rate for both the total sample and by age categories for the eligible survey sample. Age was defined as the survey participant's age at interview. The eligible survey sample includes only survey participants who were considered eligible for linkage as previously described.

Table 1 presents information for NHIS (1986-2021) by each NHIS design period. Table 2 presents information for NHANES III, NHANES 1999-2018, and 2017-March 2020 (pre-pandemic file). Linkage rates (the percentage linked out of the eligible sample) for NHIS and NHANES varied by survey years/cycles and age groups, as participants in earlier survey years and who are older are more likely to link to an NDI record.

[4] The probabilistic linkage methodology used is based on Fellegi, I. P., and Sunter, A B. (1969), "A Theory for Record Linkage," JASA 40 1183-1210.

Table 1. 1986-2021 NHIS Linked Mortality Files (with follow-up through 2022): Sample Sizes for Adults 18 and Over and Unweighted Percentages by Survey Year and Age at Interview

Survey		Total sample size	Eligible for linkage	% Eligible out of total	Linked to NDI	% Linked out of eligible
NHIS 1986-1996	Total	851,361	834,615	98.0	314,671	37.7
	18-64	708,491	694,001	98.0	186,214	26.8
	65 and over	142,870	140,614	98.4	128,457	91.4
NHIS 1997-2006	Total	687,200	613,410	89.3	140,929	23.0
	18-64	579,430	515,763	89.0	65,997	12.8
	65 and over	107,770	97,647	90.6	74,932	76.7
NHIS 2007-2018*	Total	675,715	650,047	96.2	71,955	11.1
	18-64	550,660	528,734	96.0	25,517	4.8
	65 and over	125,055	121,313	97.0	46,438	38.3
NHIS 2019-2021*	Total	93,043	91,238	98.1	3,307	3.6
	18-64	65,035	63,674	97.9	735	1.2
	65 and over	28,008	27,564	98.4	2,572	9.3

Note: Although children (<18 years) are included in the linkage, counts are not shown in the table to mitigate potential disclosure concerns.

Starting in 2007, NHIS began to collect only the last four digits of SSN.

* **Note**: For NHIS, beginning in 2015, detailed information required for linkage was collected only for the sample adult and sample child survey participants. While the NHIS collects some information about all members of the household, only sample adult and sample child participants were eligible for mortality linkage.

Table 2. NHANES III and 1999-2020 NHANES Linked Mortality Files (with follow-up through 2022): Sample Sizes for Adults 18 and Over and Unweighted Percentages by Survey Year/Cycle and Age at Interview

Survey		Total sample size	Eligible for linkage	% Eligible out of total	Linked to NDI	% Linked out of eligible
NHANES III (1988-1994)	Total	19,618	19,599	99.9	9,286	47.4
	18-64	14,366	14,350	99.9	4,226	29.4
	65 and over	5,252	5,249	99.9	5,060	96.4
NHANES 1999-2018*	Total	59,204	59,064	99.8	12,041	20.4
	18-64	45,153	45,038	99.7	3,855	8.6
	65 and over	14,051	14,026	99.8	8,186	58.4
NHANES 2017-March 2020*	Total	9,507	9,430	99.2	641	6.8
	18-64	7,124	7,060	99.1	194	2.7
	65 and over	2,383	2,370	99.5	447	18.9

Note: Although children (<18 years) are included in the linkage, counts are not shown in the table to mitigate potential disclosure concerns.

Starting in 2017-2018, NHANES included an option to provide only the last four digits of SSN.

***Note**: NHANES 2017-2018 is included in the counts for both NHANES 1999-2018 and NHANES 2017-March 2020.

4 Analytic Considerations

This section summarizes general considerations and guidelines for analysis when using the 2022 LMFs. It is not an exhaustive list of the analytic issues that researchers may encounter while using the linked data. Questions about analytic issues can be reported at datalinkage@cdc.gov.

4.1 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 required for linkage eligibility as noted in Section 3.1. 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 LMFs, analysts should limit their analysis to those survey records with a value of ELIGSTAT = 1. Linkage eligibility information for NHIS and NHANES for adults 18 and over are shown in Tables 1 and 2, respectively.

4.2 Linked Mortality File Eligibility- adjusted Sample Weights

The LMFs include sample weights adjusted for linkage eligibility (nonresponse due to insufficient identifying data) for some of the NCHS population health surveys linked to the NDI. Using standard weighting domains to reproduce population counts within these domains (sex, age, and race and ethnicity subgroups), the NCHS Data Linkage Program applied a model-based calibration approach using the SUDAAN software package (Procedure WTADJUST or WTADJX) to develop the eligibility-adjusted sample weights. Additional information on using Procedure WTADJUST to adjust sample weights for linkage eligibility, including sample SUDAAN code, is available from the NCHS Data Linkage Program.[5] [6] In addition, there are different approaches and statistical software that could be used to reduce bias to nonresponse due to insufficient identifying data.

4.2.1 NHIS Eligibility-adjusted Sampling Weights

The NCHS Data Linkage Program has provided eligibility-adjusted weights for the 1987-2021 NHIS for use with the 2022 LMFs. Treating the linkage-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.

For the 1987-2021 NHIS, participants classified as eligible for mortality follow-up had their original NHIS sampling weight adjusted to account for linkage ineligibility due to insufficient identification data. The new eligibility-adjusted sample weights provided on the 2022 LMFs are recommended for use, rather than the original NHIS sample weights, to prevent biased mortality estimates. Because there are no eligibility-adjusted sample weights for the 1986 NHIS, NCHS recommends using the public-use annual final basic weight (WTFA) for that survey year. An NCHS report assessed linkage eligibility bias for

[5] Golden, C., et al., Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare Medicaid Services. Vital Health Stat 1, 2015(58): p. 1-53.

[6] Aram, J., et al., Assessing Linkage Eligibility Bias in the National Health Interview Survey. Vital Health Stat 2, 2021(186): p. 1-28.

various sociodemographic groups and health-related variables for the 2000–2013 NHIS and supported that much of the bias was mitigated with weight adjustments.[6]

For the NHIS, the 2022 LMFs include 5 eligibility-adjusted sample weights:

(1) person-level for NHIS years 1987-2014 (FA_WGT_ADJ);
(2) sample adult in NHIS years 1997-2021 (SA_FA_WGT_ADJ);
(3) sample child in NHIS years 1997-2021 (SC_FA_WGT_ADJ);
(4) sample adult longitudinal weight for NHIS 2020, for analyses of data from 2019 and 2020 for the same individuals (SA_L_WGT_ADJ); and
(5) sample adult partial weight for NHIS 2020, for combining data from multiple years that include 2019 and 2020 (SA_P_WGT_ADJ).

The 1987-1996 NHIS did not include sample adult or sample child files, and therefore only person-level adjusted weights are provided for these years. Similarly, only sample adults and children were eligible for linkage for NHIS 2015-2021, and therefore there are no person-level adjusted weights for these years. For NHIS 2020, researchers are advised to review the NHIS 2020 Documentation for discussion of changes to field procedures in response to the COVID-19 pandemic, the three analytic data files available for sample adults, and the appropriate weight to use. The sample adult longitudinal and sample adult partial weights are only included for NHIS 2020.

4.2.2 NHANES Eligibility-adjusted Sampling Weights

The NCHS Data Linkage Program does not provide eligibility-adjusted weights for NHANES because of the survey's relatively high linkage eligibility rate (~99%). Given the very low linkage non-response, the difference between sampling weights provided on the NHANES public use files and eligibility-adjusted weights is assumed to be negligible.

4.3 Pooled Analyses of NCHS Linked Mortality Files: Pooling Survey Years/Cycles and Estimating Variance

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 and estimate variance when pooling years.

4.3.1 1992 NHIS Hispanic Oversample

The 1992 NHIS included a special oversample of the Hispanic population. The oversample was created by re-contacting Hispanic survey participants from the 1991 NHIS. Researchers planning to pool these two years of survey data should 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.[7] In addition, if researchers exclude the 1992 Hispanic oversample from pooled analyses, they should create new adjusted sample weights to properly adjust for linkage-ineligible survey participants. Guidance for the construction of new weights can be found in Appendix III of the "Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and

[7] National Center for Health Statistics, National Health Interview Survey public use data release 1992 core files - version without Hispanic oversample. 2006.

Centers for Medicare & Medicaid Services" series report.[8]

4.3.2 2017-March 2020 Pre-Pandemic NHANES

Please note that due to the suspension of the NHANES 2019-2020 field operations in March 2020, the 2019-March 2020 data were combined with the 2017-2018 data using additional weighting procedures. The resulting files are referred to as the NHANES 2017-March 2020 pre-pandemic data files. The 2017-March 2020 pre-pandemic files represent a 3.2-year period, in contrast to previous data releases which represent a 2-year period. Analysts may wish to combine 2017-March 2020 data files with previous cycles to increase sample size for outcomes with low prevalence or for subgroups. If done, the survey weights should be adjusted to reflect the longer period and larger population represented by the 2017-March 2020 files. For detailed information about analytic considerations for the NHANES 2017-March 2020 data files, refer to the NHANES website https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewbrief.aspx?Cycle=2017-2020.

4.4 Analytic Considerations for Linked NDI Data Files

4.4.1 Inconsistencies in Reported Age

Misreporting or discrepancies between reported age at interview and the date of birth may result in values for age at death that are inconsistent with baseline age when date of death and date of birth are used to calculate the age at death. The number of cases where this occurs is small, but analysts should be aware and make appropriate adjustments to the data.

4.4.2 Improbable Age Last Presumed Alive

The 2022 LMFs include records where the calculated age for participants presumed alive at the end of mortality follow-up (AGEPRALV) is 100 years or more. For these cases, there was no valid NDI record match or other source of mortality information. The NDI only includes deaths that occurred in the United States or a U.S. territory and therefore may not include death information for some deceased survey participants if they left the U.S. prior to death. 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 lost to follow-up and exclude them from the analysis.

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[9].

4.4.3 Missing Information on Date of Death

Some NDI records have missing information for the month (DODMONTH) or day of death (DODYEAR). In the 2022 LMFs, there are instances when the month or day of death are missing for survey participant records linked to the NDI. Analysts may consider imputing these values or dropping the records from their analysis.

[8] Golden, C., et al., Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare Medicaid Services. Vital Health Stat 1, 2015(58): p. 1-53.

[9] Bell, F.C. and M.L. Miller, Life Tables for the United States Social Security Area 1900-2100. 2005, Social Security Administration. https://www.ssa.gov/oact/NOTES/pdf_studies/study120.pdf

4.5 Restricted-use Linked Mortality Files Variables

4.5.1 Mortality Variables File

The linked Mortality Variables file can be used to identify which of the survey participants were eligible for NDI linkage (ELIGSTAT) and those linked to an NDI record (MORTSTAT). This file contains one record for each unique NCHS survey participant ID. Survey participant IDs with an ELIGSTAT value of 1 were considered eligible for NDI linkage.

The variable MORTSTAT indicates the vital status of a participant. Those with a MORTSTAT value of 0 are linkage eligible and assumed alive and those with a value of missing are not eligible for linkage. If a participant was linkage-eligible and considered deceased by linkage to the NDI, MORTSTAT is set to a value of 1. If the participant was linkage-eligible and considered deceased via data collection or death certification ascertainment and did not link to the NDI, MORTSTAT is set to a value of 2 (see descriptions of mortality sources below).

Although the primary determination of mortality for eligible participants is based upon matching the survey data to the NDI, additional sources of mortality information are also incorporated. These sources include data collection and ascertainment of death certificates for NCHS follow-up surveys (e.g., NHANES III). Source of mortality information is indicated by the following variables:

- MORTSRCE_NDI: 1 = mortality status ascertained through a probabilistic match to NDI record,
- MORTSRCE_DCL: 1 = mortality status obtained from the survey record (proxy decedent interview), and
- MORTSRCE_DC: 1 = mortality status obtained from a death certificate as part of survey operations.

The variable AGEDEATH provides the age at death for deceased survey data participants. For survey data participants who were not linked to an NDI record, variable AGEPRALV provides the age when the survey data participant was last presumed to be alive, which is calculated by subtracting date of birth from the end of the survey year (i.e., December 31, 2022). Underlying and multiple cause of death codes coded from the death certificate are provided for participants linked to the NDI.[10] More detailed information is available at https://www.cdc.gov/nchs/linked-data/mortality-files/index.html.

4.5.2 Death Certificate and NDI Match Variables File

The linked Death Certificate and NDI Match Variables file provides more detailed death certificate information including information on the county and state where the death occurred. This information can be used to link contextual information for analytic purposes.

Data linkages include some uncertainty over which pairs represent true matches. An estimated probability of match validity (PROBVALID) was computed for each candidate pair and compared against a probability cut-off value to determine which pairs were links (an inferred match). For additional discussion on how PROBVALID was estimated, see Appendix I – Detailed description of linkage methodology, Sections 3.3 and 3.4. NCHS used a probability cut-off value which minimized the total

[10] Less than 0.1% of the health survey participants that linked were missing cause of death information on the NDI.

estimated counts of Type I error (false positive links – identified as deceased but actually alive) and Type II error (false negative links – identified as alive but actually deceased).

In the NDI linkage with NCHS survey data, a probability cut-off value of 0.855 was used to determine final match status. Candidate pairs with a PROBVALID that exceeded the probability cut-off (i.e., PROBVALID>0.855) were considered linked. For additional discussion on probability cut-off value determination and record selection, please see Appendix I, Section 4. For some analyses, it may be desirable to reduce the Type I error. To do this, researchers should increase the probability cut-off value to a value closer to 1.0. Researchers wishing to increase the probability cut-off value should request PROBVALID in their RDC application. Note, the probability cut-off value cannot be decreased from 0.855.

In addition, individual agreement weight (pair weights components) variables are available to researchers that indicate the level of agreement among the matching variables. The total pair weight, PAIRWGT, is the sum of the partial E-M adjustment factor (see Appendix I, section 3.3) and the eight pair weight components:

- First Name or First Initial (WGT_FIRST_NAME)
- Middle Initial (WGT_MIDDLE_NAME)
- Last Name or Last Initial (WGT_LAST_NAME)
- Year of Birth (WGT_DOB_YEAR)
- Month of Birth (WGT_DOB_MONTH)
- Day of Birth (WGT_DOB_DAY)
- State of Residence (WGT_STATE_RES)
- Last 4-digits of SSN (WGT_SSN4)

Each pair weight component represents a specific identifier comparison. These component values are also available to researchers upon request in their RDC application. For more information on how the pair weight components are calculated, refer to the methodology in Appendix I, Section 3.2. When looking at the eight component pair weights simultaneously, researchers can evaluate which identifier agreements were most indicative of being a match and which identifier non-agreements were most indicative of not being a match. Mortality and death certificate variables are provided for all NCHS survey participants who linked to an NDI record (i.e., MORTSTAT=1). More detailed information is available at https://www.cdc.gov/nchs/linked-data/mortality-files/index.html.

4.6 Access to the Restricted-use Linked Mortality Files

To ensure confidentiality, NCHS provides safeguards including the removal of all personal identifiers from analytic linked files. Additionally, the linked data files are only accessible through the NCHS RDC network for approved research projects. Researchers who wish to access the restricted-use 2022 LMFs must complete an RDC application. The RDC staff will review all submitted applications to determine if the proposed project is feasible and to identify any potential disclosure risks. More information regarding the NCHS RDC network and the RDC application process is available from: https://www.cdc.gov/rdc/.

Within the RDC, the 2022 LMFs can be merged with NCHS restricted (if needed) and public use survey data files using unique survey person identification numbers (see Appendix II for merging based on PUBLICID/SEQN).

4.7 Additional Related Data Sources

Many of the NCHS surveys that have been linked to the NDI data have also been linked to HUD housing assistance program data obtained through the U.S. Department of Housing and Urban Development (HUD). Analysts interested in examining health outcomes related to housing assistance may also request variables from HUD linked files. For more information about the NCHS linked HUD files, please see the data linkage website: https://www.cdc.gov/nchs/linked-data/hud/index.html.

NCHS survey data have also been linked to Centers for Medicare & Medicaid Services (CMS) enrollment and claims data. Researchers interested in analyzing information on mortality and health care utilization for persons enrolled in Medicare may request variables from the NCHS-CMS Medicare Linkages, please see the data linkage website for more information: https://www.cdc.gov/nchs/linked-data/medicare/index.html.

Researchers interested in analyzing information on mortality and health care utilization for persons enrolled in Medicaid may request variables from the NCHS-CMS Medicaid Linkages, please see the data linkage website for more information: https://www.cdc.gov/nchs/linked-data/medicaid/index.html.

Some of the NCHS surveys linked to the NDI data have also been linked to administrative data from the Department of Veterans Affairs (VA). Researchers interested in outcomes related to Veterans may also request variables from the Linked NCHS-VA data files. The Linked NCHS-VA data files include information on a wide range of health-related topics for Veterans, including Veteran status and utilization of VA benefit programs. For more information about the linked VA data, please see the data linkage website: https://www.cdc.gov/nchs/linked-data/va/index.html.

Data users may request variables from the Linked HUD, Linked CMS Medicare, Linked CMS Medicaid, or Linked VA files in addition to the LMFs. Each of these files can be merged using the survey-specific unique participant identification variable (see Appendix II).

Appendix I: Detailed Description of Linkage Methodology

1 NHIS and NHANES and NDI Mortality Submission Files

A linkage submission file is a dataset created for conducting linkages between two sources of data. Linkage submission files, which contained the cleaned and validated PII fields, were created separately for NCHS survey data records and for NDI administrative records. The following PII fields were individually processed and output to separate files (i.e., there were separate files for SSN, DOB, name, etc., each record showing a possible value for that field for each health survey record or NDI decedent):

- SSN (validated)[11]
- DOB (month, day, and year)
- NDI DOD (month, day, and year)[12]
- Sex
- State of residence
- First, middle initial, and last name[13]

Identifier values deemed invalid by the cleaning and standardization routine were changed to a null value. A few examples where this occurred include:

- Date values: when invalid or outside of expected range
- Sex values: when multiple sex values are recorded for the same person
- Name values: multiple edits are applied:
  - Removal of special characters such as [“-.,<>/?, etc.]
  - Removal of descriptive words such as twin, brother, daughter, etc.
  - Nulling of baby names—name parts that contain specific keywords such as baby, infant, girl or boy are set to null
  - Names listed as Jane/John Doe
  - Removal of titles such as Mister, Miss, etc.
  - Removal of suffixes such as Junior, II, etc.
  - Removal of special text such as first name listed as “Void”

To increase the likelihood of finding a link, multiple or alternate submission records could be generated for each linkage eligible record in the NCHS survey data and NDI submission files based on variation of the linkage variables. Similar to the cleaning process, a more elaborate routine was used to generate alternate records involving the name fields. Alternate records were generated according to the following rules.

- Sex was missing. Two alternate records (one with male sex and the other with female) were created (note that this would result in having generated records run through both male and female specific linkage passes, and resulting duplicated links would be subsequently resolved.
- SSN with less than nine digits. A single alternate record was created where leading zeros

[11] Nine-digit SSN is considered valid if: 9-digits in length, containing only numbers, does not begin with 000, 666, or any values after 899, all 9-digits cannot be the same (i.e., 111111111, etc.), middle two and last 4-digits cannot be 0’s (i.e., xxx-00-xxxx or xxx-xx-0000), and digits are not consecutive (ex. 012345678). Additionally, special SSN values (i.e., 123-123-1234, 111-22-3333, 010-010-0101, 001-01-0001, etc.) were changed to missing. Four-digit SSN is considered valid if: 4-digits in length, containing only numbers, and is between 0001 and 9999.

[12] NDI administrative records with a missing year of death were removed from the submission file.

[13] The NDI administrative data included father’s surname which, when different from the recorded last name, was treated as an alternate last name that was used to create an alternate NDI submission record.

were added to SSN values of length 7 or 8 to make a 9-digit SSN. Note, no alternate record was created if an invalid SSN would be created by adding 0's.

- Improbable date of birth. Age at time of survey/time of death (NDI) was computed by subtracting the date of the survey (date last known alive)/date of death and the date of birth. If a date part was missing, age was computed by subtracting the year of the survey/death and the year of birth. Records with age greater than 114 had a single alternate record created,
  - If month and day were suspected of being imputed (ex. Jan 1st or June 15th), entire DOB was changed to missing[14]
  - Otherwise, only year was changed to missing
- State of residence outside of U.S. and not in rest of world (RW) list. Alternate record was created with state code changed to missing
- Multiple name parts and common nicknames (see below)

NCHS created a common nickname lookup file which was used to generate a second record replacing the nickname with the associated formal name. Similarly, multiple part names (first or last) are addressed by creating alternate name records. Table 2 below provides three examples of how alternate records were generated for nick names (Participant ID 1) and multiple part names (Participant ID 2 & 3), using hypothetical NCHS survey participant data. For participant 2, the first name was used to generate multiple records, and for participant 3, the last name was used.

Table 2. Example of Alternate Record Generation using Name Fields

Participant ID	First Name	Middle Initial	Last Name	Alternate Record
1	Beth	A	Roberts	0
1	Elizabeth	A	Roberts	1
2	Mary Ann		Davis	0
2	Mary	A	Davis	1
2	Ann		Davis	1
2	Mary		Davis	1
3	Patricia	R	Drew-Hamilton	0
3	Patricia	R	Drew	1
3	Patricia	R	Hamilton	1

NOTES: The information presented in the table was fabricated to illustrate the applied approach.

Submission files, which combined the cleaned and validated PII fields, were created separately for NCHS survey records and for NDI administrative records. During this process, multiple submission records were created for each participant/decedent to show all combinations of the recorded values for these fields. That is, if a participant/decedent had two states-of-residence recorded and three dates of birth recorded and each of the remaining fields had only one variant, then a total of six submission records would have been created for the participant/decedent (see Table 3 for example). Submission records that did not meet the eligibility requirements (see Section 3.1 Linkage Eligibility Determination) were removed from the submission file.

[14] Note, these date values are often recorded when the actual value is unknown.

Table 3. Example of Alternate Records Caused by Different PII Values

Participant ID	Day of Birth	Month of Birth	Year of Birth	State of Residence
1	31	12	1999	PA
1	30	12	1999	PA
1	15	12	1999	PA
1	31	12	1999	NY
1	30	12	1999	NY
1	15	12	1999	NY

NOTES: Data have been fabricated for this example. Other PII fields not shown as they are the same across all records.
PII – Personally Identifiable Information.

Additional post processing steps were taken after the initial survey data and NDI linkage submission files were created. First, records from both the NCHS survey data and NDI submission files were separated according to the sex value (male or female). As mentioned in section 3.2, the probabilistic linkage method assumes independence between the PII variables used to score the potential links. Records in the submission files were separated by sex to avoid violating this assumption, especially when first and/or last name and sex would be used as blocking and/or scoring variables.

2 Deterministic Linkage Using Unique Identifiers

The deterministic linkage, which was the next step in the linkage process, used only the NCHS survey data and NDI submission records that included a valid SSN. The algorithm performed two passes on the data, the first pass joining records when all 9-digits of the SSN matched and then for records where the last four digits of the SSN matched. Further, records in the 2nd pass had to have a non-missing first or last name **AND** a non-missing date of birth part (month, day, or year) to be eligible for deterministic matching using the last 4 of SSN. After records had been linked using SSN, the algorithm validated the deterministic links by comparing first name, middle initial, last name, month of birth, day of birth, year of birth, and state of residence. If the ratio of agreeing identifiers to non-missing identifiers was greater than 50% (1st pass using SSN-9) or greater than 2/3 (2nd pass using last 4 of SSN), the linked pair was retained as a deterministic match. In addition to the 2/3's agreement ratio, linked pairs in the 2nd pass were required to have at least first or last name in agreement to be deemed a deterministic match. Of note, NCHS survey data participants were excluded from the second pass (i.e., using the last 4-digits of SSN) if they were deterministically linked in the first pass. Additionally, deterministically linked records were excluded if the NCHS survey data record linked to more than 1 NDI death record or if the NDI date of death occurred more than three days before the NCHS survey date. The collection of records resulting from the deterministic match is referred to as the 'truth source.'

3 Probabilistic Linkage

The second step in the linkage process was to perform the probabilistic linkage for all records. To infer which pairs are links, the linkage algorithm first identified potential links and then evaluated their probable validity (i.e., that they represent the same individual). The following sections describe these steps in detail. The weighting procedure of this linkage process closely followed the Fellegi-Sunter paradigm, the foundational methodology used for record linkage. Based on Fellegi-Sunter, each pair was assigned an estimated probability representing the likelihood that it is a match – using pair weights computed (according to formula) for each identifier in the pair – before selecting the most probable match between two records.

3.1 Blocking

Blocking is a key step in the probabilistic record linkage process. It identifies a smaller set of potential candidate pairs, eliminating the need to compare every single pair in the full comparison space (i.e., the Cartesian product). According to Christen, blocking or indexing, “splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key).”[15] Intuitively developed rules can be used to define the blocking criteria, however, for this linkage, variable values in the data being linked were used to inform the development of a set of blocking passes that efficiently join the datasets together (i.e., multiple, overlapping blocking passes are run, each using a different blocking key). By using these data to create an efficient blocking scheme (or set of blocking passes), a high percentage of true positive links were retained while the number of false positive links were significantly reduced. A supervised machine learning algorithm used the ‘truth source’ (see Appendix I section 2) as the validation dataset and the NCHS survey data and NDI submission records as training data. For more detailed information on the supervised machine learning algorithm used, please refer to “Learning Blocking Schemes for Record Linkage” and “Using supervised machine learning to identify efficient blocking schemes for record linkage”.[16] [17]

The machine learning algorithm produced 12 blocking passes to be used in the blocking scheme. Table 4 provides the PII variables that were assigned to each of the blocking passes and the PII variables that were used to score the potential links in each of the blocking passes. Note, the variables listed in the scoring key are all PII variables not used as blocking variables.

[15] Christen, Peter. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Berlin Heidelberg: Springer-Verlag, 2012. http://www.springer.com/us/book/9783642311635.

[16] Michelson, Matthew, and Craig A. Knoblock. “Learning Blocking Schemes for Record Linkage.” In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, 440–445. AAAI’06. Boston, Massachusetts: AAAI Press, 2006. https://pdfs.semanticscholar.org/18ee/d721845dd876c769c1fd2d967c04f3a6eeaa.pdf.

[17] Campbell, S. R., Resnick, D. M., Cox, C. S., & Mirel, L. B. (2021). Using supervised machine learning to identify efficient blocking schemes for record linkage. Statistical Journal of the IAOS, 37(2), 673–680. https://doi.org/10.3233/SJI-200779.

Table 4. Blocking and scoring scheme used to identify and score potential links

Blocking Pass	Blocking Scheme	Scoring Variables
1	Month of birth, day of birth, year of birth, state of residence	First name, middle initial, last name
2	First name, month of birth, day of birth, year of birth	Middle initial, last name, state of residence
3	First name, middle initial, month of birth, state of residence	Last name, day of birth, year of birth
4	First name, month of birth, year of birth, state of residence	Middle initial, last name, day of birth
5	Last name, month of birth, year of birth	First name, middle initial, day of birth, state of residence
6	First name, day of birth, month of birth, state of residence	Middle initial, last name, year of birth
7	First name, day of birth, year of birth, state of residence	Middle initial, last name, month of birth
8	Middle initial, month of birth, day of birth, year of birth	First, Last, State
9	Last name, day of birth, state of residence	First name, middle initial, month of birth, year of birth
10	First name, middle initial, month of birth, day of birth	Last name, year of birth, state of residence
11	First name, middle initial, year of birth, state of residence	Last name, day of birth, month of birth
12	First name, last name, state of residence	Middle initial, day of birth, month of birth, year of birth

3.2 Score Pairs

Next, each pair within a given block was scored using an approach based on the Fellegi-Sunter paradigm. The Fellegi-Sunter paradigm specifies the functional relationship between agreement probabilities and agreement/non-agreement weights for each identifier used in the linkage process. The scores – pair weights – calculated in this step were used in a probability model (explained in Section 3.3), which allowed the linkage algorithm to select final links to include in the linked file. The scoring process followed the order below:

1. Calculate M- and U- probabilities (defined in Section 3.2.1)
2. Calculate agreement and non-agreement weights
3. Calculate pair weight scores

The pair scores were calculated on the agreement statuses of the following identifiers (excluding specifically the variables used to define each block – e.g., if blocking is by first name and last name, then neither were used to evaluate the pairs generated by the block):

- First Name or First Initial (when applicable)
- Middle Initial
- Last Name or Last Initial (when applicable)
- Year of Birth
- Month of Birth
- Day of Birth
- State of Residence

3.2.1 M and U Probabilities

The M-probability is the probability that the identifiers on a pair of records agree, given that records represent the same person (i.e., the records are a match). M-probabilities were estimated separately within each individual blocking pass and were calculated for each of the identifiers used for scoring (Table 4). Within the blocking pass, pairs with agreeing SSN were used to calculate the M-probabilities, as these are assumed to represent the same individual. SSN agreement was defined as having 8 or more digits being the same for pairs with a full 9-digit SSN or the last 4-digits being the same for pairs with only a 4-digit SSN (ex. XXXXX9999). Further, to account for the alternate submission records generated during the creation of the submission files, the "best" agreement was taken for each of the scoring variables among the blocked records for each NCHS survey data record ID and NDI ID (see Tables 5 and 6 for example of alternate record summarization). Table 5 is an example of how the agreement flags for each of the scoring variables in Blocking pass 1 are created. A value of 1 means the information in the variable is exactly matching, while a 0 means they are not. Table 6 then represents how the multiple submission records in Table 5 are summarized into one record for each NCHS survey data record ID and NDI administrative ID. If any of the identifiers agree across multiple records, they are flagged as agree (i.e., set to 1). The summarized records in Table 6 are then used to estimate the M-probabilities for each of the specific scoring variables.

Table 5. Example of Agreement Flags Using Blocking Pass 1

Person Identifiers		PII Agreement flags[1]		
Survey Participant ID	NDI Key	Middle Initial	Last Name	State of residence
1	1	1	0	.
1	1	.	1	0
1	1	1	0	0
2	2	1	0	0
3	789	1	1	1
3	789	0	1	1
3	789	.	1	.
3	789	0	0	1
3	322	1	0	1

NOTES: Data have been fabricated for the purposes of this example. PII – Personally Identifiable Information.

[1] Agreement status of 1 = match, 0 = non-match, and . = missing values

Table 6. Example Showing Summarization of Blocked Record Pairs for M-Probability Estimation, based on Table 5 example

Person Identifiers		PII Agreement flags[1]		
Survey Participant ID	NDI Key	Middle Initial	Last Name	State of residence
1	1	1	1	0
2	2	1	0	0
3	789	1	1	1
3	322	1	0	1

NOTES: Data have been fabricated for the purposes of this example. PII – Personally Identifiable Information.

[1] Agreement status of 1 = match, 0 = non-match, and . = missing values

Several additional comparison measures were created for first and last name identifiers in the calculation of M-probabilities:

- First/last initial agreement – used in the scoring process when only an initial was present in one or more of values (i.e., one from each of the two records being compared for a specific name variable
- Jaro-Winkler Similarity Levels – this process is explained in greater detail in Section 3.2.2

The U-probability is the probability that the two values for an identifier from paired records agreed given that they were NOT a match. Similar to the M-probabilities, U-probabilities were calculated only for the PII variables not included in the blocking keys and with the exception of first and last names, were computed within the blocking pass. The U-probabilities were computed using records where non-missing SSNs were not in agreement (defined as having less than 5 matching digits when records had a full 9-digit SSN and less than 4 matching digits for records with a 4-digit SSN). In order to avoid skewing U-probabilities in blocking passes that contained a high percentage of deterministic matches, assumed matches (i.e., records where SSN was not in agreement and had majority of the non-missing PII among scoring variables in agreement) were excluded prior to calculating the U-probabilities. For example, when computing the U-probability for day of birth in blocking pass 3, record pairs that did not agree on SSN that had a majority (i.e., greater than 50%) of the PII among first name, middle initial, month of birth, and state of residence in agreement were excluded from the assumed non-matches. Even though SSN did not agree, these records were assumed to be probable links given that a majority of the PII between the NCHS survey data records and NDI submission records agreed.

Unlike the M-probabilities, individual U-probabilities were calculated for each value of an identifier if the value was sufficiently represented in the blocking pass. Sufficient representation was defined as satisfying the following criteria:

1. Appeared in more than 2,500 record pairings (i.e., n>2,500).
2. More than 5 record pairings agreed on the value (i.e., number agree>5).
3. Agreement rate (i.e., Number or pairing that agree on value/total records pairings for that value) exceed the 5th percentile of the agreement rate across all values that met the first two conditions.

For example, if for blocking pass 1, the state of residence code for FL appeared in 30,000 record pairings, agreed on 1,560 of those pairs, and the agreement rate for state of residence exceeded the 5th percentile, then the U-probability for Florida would have been computed as 1,560/30,000=0.052 or 5.2%. A ‘catch-all’ category was created for all identifier values that did not meet the above criteria. The U-probability of the ‘catch-all’ category was computed by dividing the total number of record pairs that agreed by the total number of record pairs being used to estimate the ‘catch-all’ category. Further, if there was no agreement in the ‘catch-all’ category, the U-probability would have been set to 0. To avoid a U-probability of 0, the ‘catch-all’ U-probability was computed by halving the minimum (i.e., lowest) U-probability among the individual value’s U-probabilities. Further, if no individual value received a U-probability (i.e., all values assigned to ‘catch-all’) and there was no agreement, then the U-probability was set to 0.0001. For example, if the minimum U-probability among state of residence codes was 0.052 and there was no agreement among the catch-all records, the catch-all U-probability for state of residence would be 0.026 (0.052/2). If no state of residence code received a U-probability and there was no agreement, the U-probability for state of residence code would be 0.0001. The process for calculating U-probabilities for first and last name differs from these methods and is described in Section 3.2.2.

Lastly, an adjustment was made to the final U-probabilities to account for alternate records in the submission file. With the addition of each alternate record, the chance of agreement between the NCHS survey data records and NDI submission records increases. For example, a NCHS survey participant with different months of birth reported on participant records, has twice the chance of linking to an NDI submission record. Therefore, the U-probability for that participant’s month of birth should represent the combined chance of agreement across both month values. Section 3.2.3 provides a detailed description of the methods used to adjust the U-probabilities to account for the additional alternate submission records.

3.2.2 M and U Probabilities for First and Last Names

For first and last name M and U-probabilities, corresponding Jaro-Winkler levels (0.85, 0.90, 0.95, and 1.00) were calculated. Because agreement levels fall over a range, first and last name U-probabilities were computed for each Jaro-Winkler score level. The Jaro-Winkler algorithm assigns a string similarity score, between 0 and 1 (both inclusive), depending on the likeness between two strings. For example, if the first name on the health survey data record was “Albert” and on the NDI record it was “Abert”, this comparison would receive a Jaro-Winkler score of 0.96. M-probabilities are computed as the rate of agreement for all first/last names within a specific Jaro-Winkler level. For example, the M-probability for first name at the Jaro-Winkler 0.90 level is the rate of agreement for all first names with a Jaro-Winkler score of 0.90 and above.

Because of the large number of unique name values, it was impractical to compute specific name U-probabilities for each blocking pass (i.e., there would not be enough records available for it to be done accurately). Instead, U-probabilities were estimated using pairs generated by the Cartesian product of all records in the NCHS survey data linkage submission file and a simple random sample of 5% of records with non-missing name information from the NDI submission file (see Table 7 for the number of sampled NDI submission records).

Table 7. Count of Records from a 5% Simple Random Sample of NDI Records used to Estimate U-Probabilities for First and Last Names by Sex

	Count of Sampled Records by Name	
Sex	**First Name**	**Last Name**
Female	4,537,456	4,548,685
Male	2,530,496	2,548,147

Complete name tallies (separately, for first and last names) were then produced for the NCHS survey data linkage submission file. For each level of name on the file, 100,000 names were randomly selected from the NDI submission file 5% sample for comparison. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. For each Jaro-Winkler level, the number of names in agreement of the 100,000 randomly selected NDI file names were then tallied.[18] [19] [20]

3.2.3 Adjustment of U-Probabilities for Alternate Submission Records

As previously mentioned in section 3.2.1, an adjustment was made to the U-probabilities to account for alternate submission records. The addition of unique values for an identifier increases the likelihood of a spurious linkage between records from the files being linked. Thus, the U-probabilities were adjusted to account for the increased probability of variable agreement (i.e., if records for the same person had multiple values for a variable, the chance of agreement with any compared record from the other file increases). Therefore, NCHS survey participants received an adjusted U-probability if they had identifier values that were different across their set of submission records. The adjusted U-probabilities were then applied to each record in the set of submission records that paired with an NDI administrative record. Lastly, the U-probability that is used to compute the agreement and disagreement weights (see Section 3.2.4) is the maximum between the original and adjusted U-probability (i.e., U_{Max}=Max($U_{Original}$, U_{Adjust})).

[18] Jaro M. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J Am Stat Assoc. 1987 Jan 01;406:414-420.

[19] Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association. 1990. 354-9.

[20] Resnick, D., Mirel, L., Roemer, M., & Campbell, S. (2020). Adjusting Record Linkage Match Weights to Partial Levels of String Agreement. *Everyone Counts: Data for the Public Good*. Joint Statistical Meetings (JSM). https://ww2.amstat.org/meetings/jsm/2020/onlineprogram/AbstractDetails.cfm?abstractid=312203 (accessed August 2024).

Excluding first and last name, the adjustment process began by identifying the unique set of values, and their U-probabilities, for each of the identifiers appearing in the scoring key (Table 4), for each survey participant. Because each value is assumed to be independent of the others, the adjusted U-probabilities were computed using the additive rule for probability as the summation of the individual value U-probabilities for each participant. That is, if a NCHS survey participant had three different month of birth values, the adjusted U-probability for month of birth was simply the summation of the three individual U-probabilities. Table 8 provides an example of the process used to compute the adjusted and maximum U-probabilities for month of birth.

Table 8. Example Showing Computation of the Adjusted and Maximum U-probability for Month of Birth

Participant ID	Month of Birth	U-Probability	Adjusted U-Probability[1]	Maximum U-Probability[2]
1	6	0.091	0.253	0.253
1	5	0.083		0.253
1	7	0.079		0.253
2	1	0.110	0.191	0.191
2	10	0.081		0.191
3	6	0.091	0.091	0.091

NOTES: Data have been fabricated for the purposes of this example

[1] The adjusted U-probability is computed by summing the individual month of birth U-probabilities by participant ID.

[2] The maximum U-probability is the max U-probability value between the original and adjusted U-probabilities.

The first three columns of Table 8 show the unique values of month of birth and their corresponding U-probabilities (see Section 3.2.1) for participants 1, 2, and 3. The column titled “Adjusted U-Probability” is computed by totaling the individual probabilities in the third column for each participant. Finally, the maximum U-probability (last column), which was used to compute the agreement and disagreement weights (see Section 3.2.4), is the maximum value between the original and adjusted U-probability values.

For first and last names, only the 85% Jaro-Winkler level U-probability was adjusted. The higher levels (i.e., 90, 95, and 100) were not adjusted because of the hierarchical method being used to compute each of the U-probabilities at those levels (i.e., 90 is dependent on 85, 95 is dependent on 90, and 100 is dependent on 95). Before the 85% level was adjusted, names that were similar to one another were combined into a single name field. This step is necessary to avoid ‘double counting’ names that are highly likely to match to the same name on the NDI administrative data file. Similarity in names was defined as having a Jaro-Winkler score between 0.95 and 1 (not inclusive at the upper bound) or if one name is fully contained within another (ex. Elizabeth and Eliza). If for example, a participant had two different names, Elizabeth and Elizabith (JW_{Score}=0.967), only one would be used to adjust the 85% Jaro-Winkler U-probability. The name that is selected was determined by whichever had the highest 100% Jaro-Winkler U-probability. Using the list of ‘unduplicated’ names, the adjusted U-probability for the 85% Jaro-Winkler level was computed as the summation of each of the individual U-probabilities for the participant. Table 9 provides an example of the methods used to compute the adjusted U-probabilities for the 85% Jaro-Winkler level, using first name as an example.

Table 9. Example Showing Computation of the Adjusted and Maximum U-probability for First Name

Participant ID	First Name	U-Probability at 85% JW	U-Probability at 100% JW	Collapsed U-Probability[1]	Adjusted U-Probability[2]	Maximum U-Probability[3]
8	Margaret	0.008	0.99	0.008	0.009	0.009
8	Peggy	0.001	0.97	0.001		0.009
8	Marg	0.001	0.85	Collapsed		0.009
25	Elizabeth	0.09	0.99	0.09	0.09	0.09
25	Beth	0.01	0.95	Collapsed		0.09
78	Cathy	0.05	0.99	0.05	0.05	0.05

NOTES: Data have been fabricated for the purposes of this example. JW is the Jaro-Winkler string comparator function.

[1] The collapsed U-probability includes only the U-probabilities after similar names have been collapsed into a single name.

[2] The adjusted U-probability is computed by summing each of the collapsed 85% JW U-probabilities within each participant ID.

[3] The Maximum U-probability is the max U-probability value between the original and adjusted 85% U-probabilities.

The first four columns of Table 9 provide example Participant IDs, first names, and their U-Probabilities at the Jaro-Winkler 85 and 100 level for health survey data respondents. The collapsed U-probability column (i.e., 5th column) shows that two names were collapsed into another, i.e., for participant 8, Marg was collapsed into Margaret (full-containment) and Beth was collapsed into Elizabeth (full-containment) for participant 25. Further, the collapsed U-probability is equal to the 85% JW U-probability for the name with the highest 100% JW U-probability among the names being collapsed. The adjusted U-probability (i.e., column 6) is the summation of each collapsed U-probability for each participant ID. Finally, the maximum U-probability (i.e., last column) is the max value between the adjusted U-probability and original U-probability at the 85% JW level.

3.2.4 Calculate Agreement and Non-Agreement Weights

The agreement and non-agreement weights for each record's indicators were computed using their respective M- and U-probabilities:

$$\text{Agreement Weight (Identifier)} = \log_2\left(\frac{\text{M}}{U_{Max}}\right)$$

$$\text{Non-Agreement Weight (Identifier)} = \log_2\left(\frac{(1\text{- M})}{(1\text{ - }U_{Max})}\right)$$

Agreement weights were only assigned to identifiers that had agreeing values. Similarly, non-agreement weights were only assigned to identifiers that had non-agreeing values. A non-agreement weight was always a negative value and reduced the pair weight score. It is important to note that if the M-probability was smaller than the U-probability (i.e., M<U), the pair score (see Section 3.2.5) was not adjusted according to the agreement/non-agreement weight. Because of the logarithmic function, having a M-probability that is smaller than the U-probability would have an inverse effect on the identifier agreement weights. That is, an agreement weight computed using a M-probability that was smaller than the U-probability would produce a negative weight, while the non-agreement weight would be positive. For example, if the M-probability for month of birth was 0.989 and the U-probability was 0.9999 then the agreement and non-agreement weights would be as follows,

$$\text{Agreement Weight (Identifier)} = \log_2\left(\frac{\text{M}}{\text{U}}\right) = \log_2\left(\frac{0.989}{0.9999}\right) = -0.0158$$

Non-Agreement Weight (Identifier) = $\log_2\left(\frac{(1\text{- M})}{(1\text{ - U})}\right) = \log_2\left(\frac{0.011}{0.0001}\right)$ = 6.781

3.2.5 Calculate Pair Weight Scores

In the next step, pair weights were calculated for each record in the blocking pass, which were then used in the probability model. The pair weights were calculated differently for each blocking pass (due to different PII variables contributing to the pair weight), but followed the same general process:

1. Start with a pair weight of 0.
2. Identifier agrees: add identifier-specific agreement weight into pair weight
3. Identifier disagrees: add identifier-specific non-agreement weight (which has a negative value) into pair weight
4. Identifiers cannot be compared because one or both identifiers from the respective records compared were missing, or M-probability was less than the U-probability: no adjustment made to the pair weight

First name and last name weights were assigned using Jaro-Winkler similarity scores described in Section 3.2.2. These scores ranged from 0 to 1, with 0 representing no similarity and 1 representing exact agreement. The weighting algorithm assigned all similarity scores 0.85 and below 0.85 a disagreement weight. The algorithm assigned all similarity scores above 0.85 an agreement weight associated with the 0.85 level. If there was an agreement at the 0.85 level, the algorithm assessed the pair at the 0.90 level given that it agreed at the 0.85 level. If the names disagreed at this level, the algorithm assigned them a disagreement weight (specific to the 0.90 level given agreement at the 0.85 level). If the names agreed, the algorithm assigned them an additional agreement weight (specific to the 0.90 level). This process continued two more times: for the 0.95 and 1.00 thresholds.

3.3 Probability Modeling

A probability model, developed from a partial expectation-maximization (E-M) analysis, was applied individually to each of the blocks in the blocking scheme. Each model estimated a link probability, $P_{EM}(Match)$, for the potential matches in each blocking pass. The match probability represents the approximate likelihood that a given link is a match. These probabilities in turn allowed the linkage algorithm to:

- Combine pairs across blocking passes (Pair-weights are specific to each blocking pass and are not comparable)
- Select a “best” record among 2021 health survey data records that have linked to multiple administrative records.
- Select final matches based on a probability cut-off value (discussed in the following Section 4)

The partial E-M model was an iterative process that can be described in 4 steps:

1. A pair-weight adjustment was computed (Adj_B) specific to blocking pass, B, by taking the log base 2 of the estimated number of matches (within blocking pass B) divided by the estimated number of non-matches in the blocking pass. For convenience, the estimated number of matches, $N_{\widehat{matches,B}}$, used in the first iteration was set to half of the pairs in the blocking pass (i.e., all pairs generated by the blocking pass specification). The number of

non-matches was computed by subtracting the estimated number of matches from the number of pairs (regardless of how likely they are to be matches) in the blocking pass.

$$Adj_B = log_2\left(\frac{\widehat{N_{matches,B}}}{\widehat{N_{non-matches,B}}}\right) = log_2\left(\frac{\widehat{N_{matches,B}}}{N_{Pairs,B} - \widehat{N_{matches,B}}}\right)$$

Note that in the first iteration, it was assumed that $\widehat{N_{matches,B}}$ = $\widehat{N_{non-matches,B}}$, resulting in $Adj_B = 0$. If, however, in a later iteration, the number of matches was estimated to be, $\widehat{N_{matches,B}}$ = 20,000 (for example), out of the number of pairs, $N_{Pairs,B}$ = 1,000,000, then

$$Adj_B \ = log_2\left(\frac{20{,}000}{1{,}000{,}000 - 20{,}000}\right) \approx -5.61$$

2. The odds of a given pair, *P*, being a match were computed in blocking pass, *B*, by taking 2 to the power of the adjusted pair-weight (sum of pair-weight (*PW*) and Adj_B, the blocking pass pair weight adjustment).

$$Odds_{P,B} = 2^{PW_{P,B}+Adj,B}$$

Continuing with the example from Step 1…
if for Pair 1 of blocking pass B, the pair-weight is 8.4, then $Odds_{1,B} = 2^{(8.4+\,-5.61)}$ ≈6.9
if for Pair 2 of blocking pass B, the pair-weight is -2.5, then $Odds_{2,B} =$
$2^{(-2.5+\,-5.61)}$ ≈0.0036
…and this continues for the remaining $N_{Pairs,B}$ pairs of the blocking pass

3. Each record pair had a match probability estimated using the odds. This was accomplished by taking the odds for pair, P, in blocking pass, B, and dividing by the (Odds+1).

$$P_{EM,P,B}(Match) = \left(\frac{Odds_{P,B}}{Odds_{P,B} + 1}\right)$$

Continuing with the example…
For Pair 1 in blocking pass B, $P_{EM,P,B}(Match) = \left(\frac{6.9}{6.9+1}\right) \approx 0.87$
For Pair 2 in blocking pass B, $P_{EM,P,B}(Match) = (\frac{0.0036}{0.0036+1}) \approx 0.0036$
…and this continues for the remaining $N_{Pairs,B}$ pairs of the blocking pass.

4. The new number of matches in blocking pass were estimated. This was done by summing each of the estimated probabilities in the block.

$$\widehat{N_{matches,B}} = \sum \widehat{P_{EM,P,B}(Match)}$$

Continuing with the example, add the probabilities for every pair in the blocking pass:

$$\widehat{N_{matches,B}} = 0.87 + .0036 + \widehat{P_{EM,3,B}} + \ldots + \widehat{P_{EM,N_{Pairs,B},B}}$$

This process was repeated until convergence was reached in the number of matches being estimated. Once convergence was achieved, the final probabilities were estimated based on the last value of $\widehat{N_{matches,B}}$ to be estimated. These estimated probabilities were then used to select the final matches, as described below in Section 4.

3.4 Adjustment for SSN Agreement

Up to this point, every pair generated through the probabilistic routine was assigned a value that estimates its probability of being a match. However, this estimate did not take SSN agreement into account. This was conducted as a separate step because for the other comparison variables, M- and U-probabilities were estimated based on probable matches or non-matches that were determined based on SSN agreement, and clearly this was infeasible for SSN itself.[21]

To remedy this, before the algorithm adjudicated the matches against the probability cut-off value, one final adjustment was made to the match probabilities (for probabilistic pairs). For pairs that had an SSN on both the NCHS survey data record and NDI submission record, the estimated probability was adjusted based on the last four digits of the SSN.

When the last four digits of SSN agreed (i.e., are exactly the same):

$$Probvalid_{SSN_{Adj}} = \frac{\left(\frac{P_{EM}(Match)}{1-P_{EM}(Match)} \cdot \frac{M_{SSN-SSN4}}{U_{SSN-SSN4}}\right)}{\left(\left(\frac{P_{EM}(Match)}{1-P_{EM}(Match)} \cdot \frac{M_{SSN-SSN4}}{U_{SSN-SSN4}}\right)+1\right)}$$

When the last four digits of SSN did not agree:

$$Probvalid_{SSN_{Adj}} = \frac{\left(\frac{P_{EM}(Match)}{1-P_{EM}(Match)} \cdot \frac{(1-M_{SSN-SSN4})}{(1-U_{SSN-SSN4})}\right)}{\left(\left(\frac{P_{EM}(Match)}{1-P_{EM}(Match)} \cdot \frac{(1-M_{SSN-SSN4})}{(1-U_{SSN-SSN4})}\right)+1\right)}$$

No adjustment was made for pairs that did not have an SSN on either the NCHS survey data record or NDI submission record. So, for these pairs:

$$Probvalid_{SSN_{Adj}} = P_{EM}(Match)$$

4 Estimate Linkage Error, Set Probability Cut-off Value, and Select Matches

The scored (probabilistic) and deterministic linkage files for males and females were combined prior to estimating the linkage error and selecting matches. Recall the purpose for separating the records by sex was to avoid violating the independence assumption for name identifiers mentioned by Fellegi-Sunter. Now that records from each sex have been separately scored, there is no need to keep them separate.

[21] The M and U probabilities in the formulas refer specifically to the M and U of the last four digits of the SSN.

4.1 Estimating Linkage Error to Determine Probability Cut-off Value

Subsequent to performing the record linkage analysis an error analysis was performed. There are two type of errors that were estimated:

- Type I Error: Among pairs that are linked, what percentage of them were not true matches.
- Type II Error: Among true matches, how many were not linked.

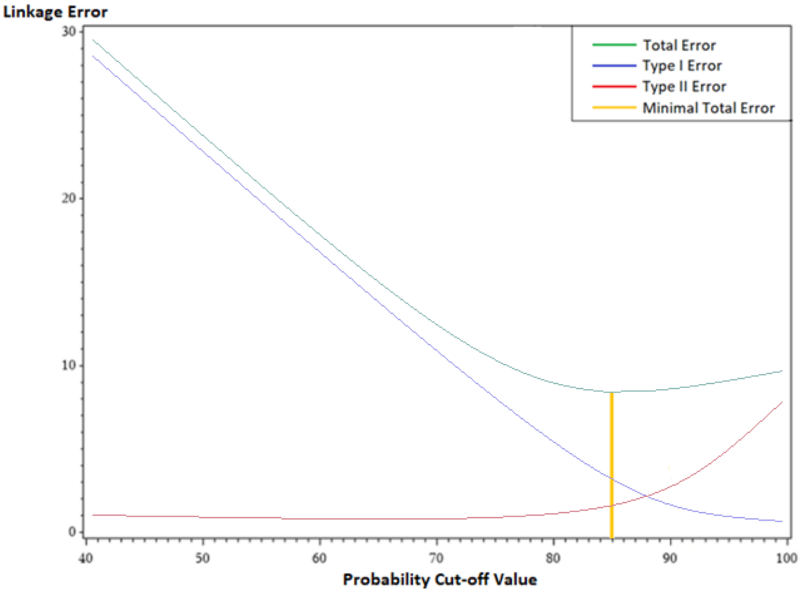
Because all records were included in the probabilistic linkage (i.e., even deterministic links), SSN agreement status (defined as seven or more matching digits for nine-digit SSN's and for SSN's that had only the last four digits, all four digits must match) was used to measure Type I error. Type I error for probabilistic links was measured as the total number of probabilistic links with non-agreeing SSN divided by the total number of probabilistic links with a valid SSN available on both the NCHS survey data record and NDI submission record. Also, deterministically established links were considered to have 0% Type I error rates. While it was believed that the error for these links was quite small and near 0, it is expected that some error does exist even with the deterministically established links and so the estimate was likely biased low. For example, if 40% of links were derived from the deterministic method, this would reduce the estimated Type I error by the proportion of probabilistically determined linkages among all linkages. To further illustrate, if the Type I error rate for probabilistic links was estimated as 1.2%, then the estimated Type I error rate for the combined linkage process would be (0.40*0.012) = 0.0048 or 0.48%.

To measure Type II error, the truth source comprised of all matches identified in the deterministic linkage was used. Recall, the truth source contains records with full nine-digit SSN agreement (step 1) or with the last four digits of SSN in agreement (step 2). Potential deterministic matches were then validated using the available PII (see, Appendix I section 2). It was expected that this truth source had only a few exceptional pairs that were not true matches. For the probabilistic records, Type II error was estimated as the percentage of the truth source records that were not returned as links by the probabilistic method. Similarly to the computation of Type I error, an adjustment was made to the Type II error since some links having agreeing SSNs were being linked deterministically even if they were not returned by the probabilistic approach. For example, say that the probabilistic approach was able to return 97% of true matches as links. If only a probabilistic linkage was conducted, the Type II error would then be 3%. However, among the 3% not linked probabilistically, some pairs could be linked deterministically. If the deterministic linkage rate is 50% (and if we assume the same rate among the non-linked pairs), then the Type II error rate can be estimated as 0.5*(1 – 0.97) = 0.015 or 1.5%.

4.2 Set Probability Cut-off Value

One goal of record linkage is to have the lowest errors possible. However, as more pairs are accepted, pairs that are less certain to be matches but accepted as links increase the Type I error and decrease Type II error. And as less pairs are accepted, pairs that are more certain to be matches but not accepted as links decrease the Type I error and increase Type II error. The optimal trade-off between Type I error and Type II error is not known, but it can be assumed to be optimal when the sum of Type I and Type II error is at a minimum. For this reason, Type I and Type II error are estimated at various probability cut-off values and the one that showed the lowest estimate of total error is selected (see Figure 1). For the linkage of the NCHS survey data and NDI, the optimal probability cut-off value was set to 0.855.

Figure 1. Illustrating linkage error by probability cut-off value
(Illustrative schematic not based on actual values)



4.3 Select Links Using Probability Cut-off Value

The final step in the linkage algorithm was to determine links, which were record pairs inferred to be matches. Links were pairs where the $Probvalid_{SSN_{Adj}}$ exceeded the probability cut-off value (from Section 4.2). Further, the 'best' record pair (i.e., highest $Probvalid_{SSN_{Adj}}$) among the records that exceeded the probability cut-off value was selected for each survey participant. Additionally, records were excluded from the set of selected links if one of the following was true,

- Full NDI date of death (i.e., non-missing day, month, and year). If the NDI date of death occurred more than 3-days prior to the interview date on the NCHS survey data record.
- Partial NDI date of death (i.e., either day or month were missing).
    - Month of death is known, and day of death is unknown. Month and year of death must occur on or after the month and year of the interview date on the NCHS survey data record.
    - Month of death is unknown. Year of death must occur on or after the year of the interview date on the NCHS survey data record.

If any of the above conditions were true, the record was excluded. All record pairs with an adjusted probability value that fell below the probability cut-off value were not linked.

4.4 Computed Error Rates of Selected Links

Final error rates were computed for selected links (described in Section 4.3). Table 10 provides the total number of selected links, the number of total links identified through deterministic and probabilistic methods, and the Type I and Type II error rates for the NCHS survey – NDI linkage. Because the links were selected using the SSN adjusted probability (described in Section 4.1), the overall Type I error rate was computed using the estimated match probabilities rather than using SSN agreement. For the probabilistic links, the estimated match probabilities represented the probability that the NCHS survey data record was a match to the NDI administrative record. In other words, if a link had an estimated probability of 0.98, then it was understood that there was a 98% chance this was a match. To estimate the Type I error rate for the probabilistic links, the chance that a link is not a match was summed (i.e., $\sum 1 - Probvalid_{SSN_{Adj}}$) and then divided by the total number of probabilistic records. The method to measure the overall Type II error remained unchanged (see Section 4.1).

Table 10. Algorithm Results for Total Selected Links

	Probability Cut-off Value	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
NCHS Surveys	0.855	576,837	377,476	199,361	0.09%	1.48%

Appendix II: Merging Restricted-use LMF Data and Public-use NCHS Survey Data

The data provided on the 1986-2021 NHIS, 1999-2018 NHANES and 2017-March 2020 NHANES pre-pandemic files, as well as NHANES III LMFs can be merged with the NCHS restricted and public use survey data files using the unique survey specific public identification number (PUBLICID/SEQN).

Note: The 2022 LMFs are only available for research use through the NCHS RDC Network. Approved RDC researchers may choose to provide their own analytic files created from public use survey files to the RDC. Therefore, it is important for researchers to include the survey specific public identification number on any analytic files sent to the RDC. The RDC will merge data (using PUBLICID or SEQN) from the 2022 LMF to the analyst's file. The merged file will be held at the RDC and made available for analysis.

Information on how to identify and/or construct the NCHS survey specific PUBLICID or SEQN is provided below.

1 National Health Interview Survey (NHIS), 1986-2021

1.1 NHIS 1986-1994

	Public-use		
Variable	Location	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 within quarter
PNUM	15-16	2	Person number within household

Concatenate all variables to get the unique person identifier.

SAS example:

```
length PUBLICID $14;
PUBLICID = ((put(YEAR,2.)) || (put(QUARTER,1.)) || PSUNUMR || (put(WEEKCEN,z2.)) || SEGNUM || HHNUM ||
PNUM);
```

Stata example:

```
egen PUBLICID = concat(YEAR QUARTER PSUNUMR WEEKCEN SEGNUM HHNUM PNUM) (Note that this
will convert the variables to string variables.)
```

R example:

```
# Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%02d", df$YEAR), sprintf("%01d", df$QUARTER),sprintf("%03d",
df$PSUNUMR), sprintf(“%02d”, df$WEEKCEN), sprint(“%02d”, df$SEGNUM), sprint(“%02d”,
df$HHNUM), sprintf("%02d", df$PNUM))
```

1.2 NHIS 1995-1996

	Public-use		
Variable	Location	Length	Description
YEAR	3-4	2	Year of interview
HHID	5-14	10	Household ID number
PNUM	15-16	2	Person number within household

Concatenate all variables to get the unique person identifier.

SAS example:

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

Stata example:

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

(Note that this will convert the variables to string variables.)

R example:

```
# Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%02d", df$YEAR), sprintf("%10d", df$HHID),sprintf("%02d", df$PNUM))
```

1.3 NHIS 1997-2003

	Public-use		
Variable	Location	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

Concatenate all variables to get the unique person identifier. The person identifier was called PX in the 1997-2003 NHIS and FPX in the 2004 (and later) NHIS; users may find it necessary to create an FPX variable in the 2003 and earlier datasets (or PX in later datasets).

SAS example:

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

Stata example:

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

(Note that this will convert the variables to string variables.)

R example:

```
# Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%04d", df$SRVY_YR), sprintf("%06d", df$HHX),sprintf("%02d",
df$FMX),sprintf("%02d", df$PX))
```

1.4 NHIS 2004

Variable	Public-use Location	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

Concatenate all variables to get the unique person identifier.

SAS example:

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

Stata example:

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

(Note that this will convert the variables to string variables.)

R example:

```
# Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%04d", df$SRVY_YR), sprintf("%06d", df$HHX),sprintf("%02d",
df$FMX),sprintf("%02d", df$FPX))
```

1.5 NHIS 2005-2018

Variable	Public-use Location	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

Concatenate all variables to get the unique person identifier.

SAS example:

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

Stata example:

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

(Note that this will convert the variables to string variables.)

R example:

```
# Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%04d", df$SRVY_YR), sprintf("%06d", df$HHX),sprintf("%02d",
df$FMX),sprintf("%02d", df$FPX))
```

1.6 NHIS 2019-2021

Variable	Public-use Location	Length	Description
SRVY_YR	3-6	4	Year of interview
HHX	7-13	7	Household number
RECTYPE	1-2	2	Record type

Note: The NHIS public-use files since the 2019 redesign do not contain a person number variable. To merge multiple NHIS public-use files, follow instructions provided in NHIS documentation. To merge to the linked data, concatenate the above variables from the Sample Adult file (RECTYPE=10 for all Sample Adults) or the Sample Child file (RECTYPE=20 for all Sample Children) to get the unique person identifier.

SAS example:

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

Stata example:

```
egen PUBLICID = concat(SRVY_YR HHX RECTYPE)
```

(Note that this will convert the variables to string variables.)

R example:

```
# Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%04d", df$SRVY_YR), sprintf("%07d", df$HHX),sprintf("%02d",
df$RECTYPE))
```

2 National Health and Nutrition Examination Surveys (NHANES)

2.1 Third National Health and Nutrition Examination Survey (NHANES III)

Item	Length	Description
SEQN	5	Participant identification number

All of the NHANES III public-use data files are linked with the common survey participant identification number (SEQN). Merging information from multiple NHANES III files to the 2022 LMF using this variable ensures that the appropriate information for each survey participant is merged correctly.

2.2 National Health and Nutrition Examination Survey (NHANES), 1999-2018

Item	Length	Description
SEQN	6	Participant identification number

All of the NHANES public-use data files are linked with the common survey participant identification number (SEQN). Merging information from multiple NHANES files to the 2022 LMFs using this variable ensures that the appropriate information for each survey participant is merged correctly.

2.3 NHANES 2017-March 2020 Pre-Pandemic Data

Item	Length	Description
SEQN	6	Participant identification number

All of the NHANES public-use data files are linked with the common survey participant identification number (SEQN). Merging information from multiple NHANES files to the 2022 LMFs using this variable ensures that the appropriate information for each survey participant is merged correctly.
Note: Comparisons or examination of trends between 2017-2018 and 2019-March 2020 data are not possible and should not be conducted because the 2019-March 2020 data do not represent any defined population. To prevent data analysts from separating the two cycles, changes have been made to respondent sequence identification numbers and PSU and strata variables have been masked.