The Linkage of the National Center for Health Statistics Survey Data to United States Department of Veterans Affairs Administrative Data:

Linkage Methodology and Analytic Considerations

Data Release Date: September 16, 2022

Document Version Date: January 31, 2025

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 the National Center for Health Statistics Survey Data to United States Department of Veterans Affairs Administrative Data: Linkage Methodology and Analytic Considerations, January 2025. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data/datalinkage/NCHS-VA-Linked-Data-Methodology-and-Analytic-Considerations.pdf

Table of Contents

1 Introduction....6

2 Data Sources....7

2.1 National Health Interview Survey (NHIS)....7

2.2 National Health and Nutrition Examination Survey (NHANES)....7

2.3 VA Benefit Programs and Data....7

2.3.1 VA Administrative Data....8

3 Linkage Methodology....9

3.1 Linkage Eligibility Determination....9

3.1.1 Match Rate Tables....10

3.2 Overview of Linkage....11

3.3 Description of NCHS-VA Linked Data Files....11

3.3.1 Match Status File....12

3.3.2. Service Record File....13

3.3.3. VA Utilization File....14

4 Analytic Considerations....15

4.1 Access to the Restricted-Use NCHS-VA Linked Data Files....15

4.2 Variables to Request in RDC Proposals....15

4.3 Merging NCHS-VA Linked Data Files with NCHS Survey Data....16

4.4 2005-2018 NHANES Sample Sizes....17

4.5 Linkage Eligibility-adjusted Participant Survey Weights....17

4.6 Temporal Alignment of VA Benefit Program Utilization....17

5 Additional Related Data Sources....18

Appendix I: Detailed Description of Linkage Methodology....20

1 NCHS and VA Linkage Submission Files....20

2 Deterministic Linkage Using Unique Identifiers....22

3 Probabilistic Linkage....22

3.1 Blocking....22

3.2 Score Pairs....24

3.2.1 Calculate M- and U- Probabilities....24

3.2.2 M- and U-Probabilities for First and Last Names....26

3.2.3 Calculate Agreement and Non-Agreement Weights....27

3.2.4 Calculate Pair Weight Scores....27

3.3 Probability Modeling....28

3.4 Adjustment for SSN Agreement....29

4 Estimate Linkage Error, Set Probability Threshold, and Select Matches 30
4.1 Estimating Linkage Error to Determine Probability Cutoff 30
4.2 Set Probability Cutoff 32
4.3 Select Links Using Probability Threshold 33
4.4 Computed Error Rates of Selected Links 33
Appendix II: Merging NCHS-VA Linked Data Files with NCHS Survey Data 34

List of Acronyms

CMS, Centers for Medicare & Medicaid Services

DOB, date of birth

DoD, Department of Defense

EM, expectation-maximization

ERB, Ethics Review Board

FY, Fiscal Year

HICN, Health Insurance Claim Number

HUD, Department of Housing and Urban Development

NCHS, National Center for Health Statistics

NDI, National Death Index

NHANES, National Health and Nutrition Examination Survey

NHIS, National Health Interview Survey

PII, Personally Identifiable Information

RDC, Research Data Center

SSA, Social Security Administration

SSN, Social Security number

SSN9, 9-digit Social Security number

SSN4, Last four digits of Social Security number

USVETS, United States Veterans Eligibility Trends and Statistics

VA, Department of Veterans Affairs

VBA, Veterans Benefits Administration

VHA, Veterans Health Administration

1 Introduction

As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to collect, analyze, and disseminate timely, relevant, and accurate health data and statistics. NCHS products and services inform the public and guide program and policy decisions to improve our nation’s health. In addition to collecting and disseminating the nation’s official vital statistics, NCHS conducts several population-based surveys, including the National Health Interview Survey (NHIS), https://www.cdc.gov/nchs/nhis/index.htm (accessed September 13, 2022), and the National Health and Nutrition Examination Survey (NHANES), https://www.cdc.gov/nchs/nhanes/index.htm (accessed September 13, 2022). These surveys provide rich cross-sectional information on population characteristics and risk factors such as smoking, height and weight, health status, and socio-economic circumstances. Although the survey data collected 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 health-related administrative data sources. The linkage of survey and administrative data provides the unique opportunity to study changes in health status, health care utilization and expenditures, and other outcomes in specialized populations, such as people receiving housing assistance.

In a collaboration with the US Department of Veterans Affairs (VA), the NCHS Data Linkage Program has been able to expand the analytic utility of the data collected from NHIS and NHANES by augmenting it with administrative data collected by VA. **This report will describe the linkage of data from the 2005-2018 NHIS and the 2005-2018 NHANES to VA administrative data through September 30, 2020 (fiscal year 2020).** This linkage, collectively referred to as the NCHS-VA Linked Data Files, creates a new data resource that can support research studies focused on a wide range of health topics for Veterans, including Veteran status and utilization of VA benefit programs.

This document describes the first linkage conducted between NCHS survey data and VA administrative data. A brief overview of the data sources, a description of the methods used for linkage, description of the linked data files, and analytic considerations are included in this document to assist researchers when using the linked files. Detailed information on the linkage methodology is provided in Appendix I: Detailed Description of Linkage Methodology. More information about VA benefit programs can be found on the VA website.[1] Additional documentation about the variables in the linked data files are available from the NCHS data linkage website.[2]

The data linkage work was performed at NCHS in part through contract #HHSD2002016F92236B by NORC at the University of Chicago, with funding from the Centers for Disease Control and Prevention Data Modernization Initiative.

[1] VA. https://www.va.gov see drop down tab “VA Benefits and Health Care” (accessed September 13, 2022).

[2] NCHS. Restricted-Use NCHS-VA Data. https://www.cdc.gov/nchs/data-linkage/va-restricted.htm (accessed September 13, 2022).

2 Data Sources

2.1 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 US. 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, persons within households, and finally, one selected sample adult and sample child. It should be noted that individuals who are on active duty in the US uniformed services (non-civilians) at the time of survey sampling are considered ineligible for participation in the NHIS. The survey has been conducted continuously since 1957 and the content is periodically updated. Prior to 2007, NHIS traditionally collected full 9-digit Social Security Numbers (SSN9) from survey participants. However, in an attempt to address respondents' increasing refusal to provide SSN and implicit consent for linkage, in 2007 the NHIS began to collect only the last four digits of SSN (SSN4) and added an explicit consent question about linkage for those who refused to provide SSN. Additionally, from 2007 onward, only sample adults and children were eligible for linkage. The implications of this procedural change on data linkage activities are discussed later in this report. The NHIS sample sizes can vary from year to year and are dependent on the sample design, which changes periodically. The annual sample size (i.e., number of households) for the sample design starting in 2016, and for the previous 2006-2015 sample design is about 35,000 households, with each household having one adult selected to be the sample adult. For more information on the 1995-2004 sample design which was extended to 2005, refer to NCHS Series 2 report (Number 130).[3] Only sample adults from the 2005-2018 NHIS are included in this linkage. For detailed information on the NHIS data content and methods, refer to the NHIS website, http://www.cdc.gov/nchs/nhis.htm (accessed September 13, 2022).

2.2 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 concerns, 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 US civilian, noninstitutionalized population that is selected using a complex, multistage probability design. Similar to NHIS, individuals who are on active duty in the US uniformed services (non-civilians) at the time of survey sampling are considered not eligible for participation in NHANES. Only adults 18 and older from the 2005-2018 NHANES cycles are included in this linkage. For detailed information about the Continuous NHANES data contents and methods, refer to the NHANES website, https://www.cdc.gov/nchs/nhanes/index.htm (accessed September 13, 2022).

2.3 VA Benefit Programs and Data

The VA provides lifelong benefits to eligible military Veterans and their families.

[3] NCHS Series 2 (Number 130). Vital and Health Statistics. Design and Estimation for the National Health Interview Survey, 1995-2004. https://www.cdc.gov/nchs/products/series.htm (accessed September 13, 2022)

Benefits include VA health care administered by the Veterans Health Administration (VHA), which serves 9 million enrolled Veterans each year at nearly 1,300 integrated health care facilities.[4, 5] Eligibility for VA health care includes prior active-duty service and is dependent on factors such as the character of separation (e.g., honorable or dishonorable), timing, and length of active-duty service. Enhanced eligibility status (placement in a higher priority group, which increases the likelihood a person will be eligible for benefits) is further dependent on factors such as having a service-connected disability[6].

Through the Veterans Benefits Administration (VBA), the VA also helps service members transition out of active-duty service, and assists with service-connected disability compensation[7], pension[8], VA guaranteed home loans[9], life insurance[10], education and training[11], veteran readiness (vocational rehabilitation)[12], and other benefits.[13]

2.3.1 VA Administrative Data

VA administrative data contains Veteran-level information on active-duty in the US uniformed services (such as branch of service, time since last separation from active-duty, and era of service) and VA benefit program utilization including: VA health care, service-connected disability compensation, pension, VA guaranteed home loan program, life insurance, education, training, and veteran readiness (vocational rehabilitation), and employment benefit programs. The VA offers additional benefits and services, such as burial and memorial services, that are not included in the NCHS-VA Linked Data Files.

The VA administrative data included in the NCHS-VA Linked Data Files was extracted from the United States Veterans Eligibility Trends and Statistics (USVETS) information management system. USVETS is an integrated data source on all US Veterans (living and deceased). It is produced by the VA Office of Data Governance and Analytics, within the Office of Enterprise Integration, to support operational and policy issues throughout the VA. Examples of USVETS data sources include Department of Defense (DoD), VHA, and VBA. The USVETS dataset contains one record per Veteran, following an adjudication process that aggregates data from across different data sources.[14] USVETS provides a comprehensive picture of the Veteran population to support statistical, trend, and longitudinal analysis. Not all information is sourced directly from administrative records. For example, race/ethnicity may be supplied from purchased data sources, if no better source exists, and sex may be imputed based on name. The USVETS database may not include all Veteran records, particularly among older ages (e.g., 70 and older). Additionally, information on some Veterans who have not had a relationship with VA, and/or

[4] VA Health Care. https://www.va.gov/health-care/ (accessed September 13, 2022).

[5] VHA. https://www.va.gov/health/ (accessed September 13, 2022).

[6] VA. Eligibility for VA health care. https://www.va.gov/health-care/eligibility/ (accessed September 13, 2022).

[7] VA. Compensation. https://www.benefits.va.gov/compensation/index.asp (accessed September 13, 2022).

[8] VA. Pension. https://www.benefits.va.gov/pension/index.asp (accessed September 13, 2022).

[9] VA. VA Home Loans. https://www.benefits.va.gov/homeloans/index.asp (accessed September 13, 2022).

[10] VA. Life Insurance. https://www.benefits.va.gov/insurance/index.asp (accessed September 13, 2022).

[11] VA. VA education and training benefits. https://www.va.gov/education/ (accessed September 13, 2022).

[12] VA. Veteran Readiness and Employment (VR&E). https://www.benefits.va.gov/vocrehab/index.asp (accessed September 13, 2022).

[13] VA. Summary of VA Benefits. https://benefits.va.gov/BENEFITS/benefits-summary/SummaryofVABenefitsFlyer.pdf (accessed September 13, 2022)

[14] USVETS, Data Governance & Analytics, Office of Enterprise Integration, Department of Veterans Affairs.

whose active-duty service was prior to 1970, may not be complete. This linkage includes VA administrative data through fiscal year (FY) 2020.

Data products and reports that use USVETS data, as well as descriptions of the Veteran population, can be found at the website for the VA National Center for Veteran Analysis and Statistics.[15]

3 Linkage Methodology

3.1 Linkage Eligibility Determination

The linkage of these data was conducted through an agreement between NCHS and VA. Approval for the linkage was provided by NCHS' Research Ethics Review Board (ERB).[16] The data linkage work was performed at NCHS.

Only a subset of 2005-2018 NHIS and 2005-2018 NHANES participants were eligible for linkage with the VA administrative data. NCHS survey participants who have provided consent as well as the necessary personally identifiable information (PII), such as name and date of birth (DOB), are considered linkage eligible. Linkage eligibility refers to the potential ability to link data from an NCHS survey participant to administrative data. Criteria for NCHS-VA linkage eligibility vary by survey and year due to variability of linkage related questions and changes to PII collection procedures by the surveys. Survey participants younger than 18 years of age were considered ineligible for the NCHS-VA linkage, due to small sample size concerns. Additionally, only sample adults from NHIS were considered linkage eligible for this data linkage.

For NHIS prior to 2007, a refusal by the survey participant to provide a SSN9 was considered an implicit refusal for data linkage. However, NCHS observed an increase in the refusal rate for providing SSN, particularly for NHIS, which reduced the number of survey participants eligible for linkage.[17] In an attempt to address declining linkage eligibility rates, NCHS introduced new procedures for obtaining consent for linkage from survey participants in 2007 for NHIS. Research was also conducted to assess the accuracy of matching data from NHIS to the National Death Index (NDI) using partial SSN and other PII.[18] The research assessed algorithms using the last four and last six digits of SSN. The results were favorable and provided sufficient evidence to support changes in how NHIS collected SSN for linkage.[19] Therefore beginning in 2007, NHIS started requesting only the SSN4. In addition, a short introduction before asking for SSN4 was added, and participants who declined to provide SSN4 were asked for their explicit permission to link to administrative records without SSN. Also, at this time, the NCHS ERB determined that

[15] VA. National Center for Veterans Analysis and Statistics. https://www.va.gov/vetdata/ (accessed September 13, 2022).

[16] The NCHS Research ERB, also known as an Institutional Review Board or IRB, is an administrative body of scientists and non-scientists that is established to protect the rights and welfare of human research subjects.

[17] Miller, D.M., R. Gindi, and J.D. Parker, Trends in record linkage refusal rates: Characteristics of National Health Interview Survey participants who refuse record linkage. Presented at Joint Statistical Meetings 2011. Miami, FL., July 30–August 4.

[18] Sayer, B. and Cox, C.S. How Many Digits in a Handshake? National Death Index Matching with Less Than Nine Digits of the Social Security Number in Proceedings of the American Statistical Association Joint Statistical Meetings. 2003.

[19] Dahlhamer, J.M. and Cox, C.S., Respondent Consent to Link Survey Data with Administrative Records: Results from a Split-Ballot Field Test with the 2007 National Health Interview Survey. paper presented at the 2007 Federal Committee on Statistical Methodology Research Conference, Arlington, VA, 2007.

for the 2007 NHIS and all subsequent years, only sample adult and sample child survey participants (and not other persons within household, see section 2.1) were eligible for data linkage.

Sample adult participants in the 2007-2018 NHIS were considered eligible for the VA linkage if they:

- Provided SSN4 or an affirmative response to the follow-up question to allow linkage without SSN4,
- Provided sufficient data elements for linkage, and
- Were at least 18 years old at the time of interview.

Sample adult participants in the 2005-2006 NHIS were considered eligible for the VA linkage if they:

- Did not refuse to provide SSN9,
- Provided sufficient data elements for linkage, and
- Were at least 18 years old at the time of interview.

For NHANES, the informed consent procedures changed as well. SSN9 was consistently collected across the survey cycles for 2005-2018. For NHANES cycles prior to 2009-2010, a refusal by the survey participant to provide a SSN9 was considered an implicit refusal for data linkage. However, beginning with the 2009-2010 NHANES, participants were explicitly asked for consent to be included in data linkage activities during the informed consent process prior to the interview. Only participants who provided an affirmative response to the linkage consent question were considered linkage eligible.

Participants in the 2009-2018 NHANES were considered eligible for the VA linkage if they:

- Provided an affirmative response to the linkage consent question,
- Provided sufficient data elements for linkage, and
- Were at least 18 years of age at the time of screener.

Participants in the 2005-2008 NHANES were considered eligible for the VA linkage if they:

- Did not refuse to provide SSN9,
- Provided sufficient data elements for linkage, and
- Were at least 18 years of age at the time of screener.

Note that linkage eligibility is distinct from benefit program eligibility, which defines whether a person meets the eligibility criteria for a specific VA-administered or funded program. More information about VA eligibility criteria is available from the VA website.[20]

3.1.1 Match Rate Tables

Match rate tables providing NCHS-VA linked sample sizes (the numbers and percentages of those who were eligible for linkage and the number who were linked to VA administrative data) for 2005-2018 NHIS and 2005-2018 NHANES participants are available at https://www.cdc.gov/nchs/data-linkage/va-methods.htm.

[20] VA. VA Benefits and Health Care. https://www.va.gov/ (accessed September 13, 2022).

For NHIS, match rates and sample sizes are shown for all survey years and overall, by age group (18-64 years and 65 and over) and sex. For NHANES, the match rates and sample sizes are shown by age group (18-64 years and 65 and over) and sex for all cycles combined but for each NHANES cycle, match rates and sample sizes are shown by age group only, due to disclosure concerns.

3.2 Overview of Linkage

This section outlines steps that were used to link the NCHS data to the VA administrative data. For more detailed information on linkage methodology see Appendix I: Detailed Description of Linkage Methodology.

Data from linkage-eligible NCHS participants were linked to the VA administrative records using the following identifiers: SSN, first name, last name, middle initial, month of birth, day of birth, year of birth, 5-digit ZIP code of residence, state of residence, and sex.

Data from NCHS survey participants and the VA administrative records were linked using both deterministic and probabilistic approaches. For the probabilistic approach, scoring was conducted according to the Fellegi-Sunter method.[21] Following this, a selection process was implemented with the goal of selecting pairs that represented the same individual between the data sources. The following three steps were applied to determine linked records:

1. Deterministic linkage joined records on exact SSN, with links validated 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 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). Deterministic matches (from step 1) were assigned a match probability of 1 and records selected from the probabilistic match (step 2) were assigned the modeled match probability.

For each NCHS participant record that was linked, VA extracted information from the USVETS database and sent the data to NCHS through a secure data transfer system.

3.3 Description of NCHS-VA Linked Data Files

The NCHS-VA Linked Data Files are comprised of the Match Status File, the Service Record File, and the VA Utilization File, with the latter two files derived from the USVETS data. Variables found in each file are referenced in the data dictionaries. The Service Record File includes information detailing active-duty service in the US uniformed services. The VA Utilization File includes information from the VA on enrollment, status (e.g., healthcare enrollment priority status), and utilization related to VA benefit programs.

[21] Fellegi, I. P., and Sunter, A B. (1969), "A Theory for Record Linkage," JASA 40 1183-1210.

All the files include survey identification variables – The variable SURVEY (identifying survey name) identifies the source survey for each record. The variables PUBLICID and SEQN are unique identifiers specific to each survey (NHIS and NHANES, respectively) and allow the different NCHS-VA Linked Data Files to be merged with the public-use NCHS survey data files using unique person identifiers. More information on merging files can be found in Appendix II: Merging NCHS-VA Linked Data Files with NCHS Survey Data.

3.3.1 Match Status File

The Match Status File can be used to identify which survey participants were eligible for linkage and linked to VA administrative records. In addition, the file includes linkage eligibility-adjusted weights for each survey and an indicator of linkage certainty. The Match Status File contains a record for each 2005-2018 NHIS sample adult and each 2005-2018 NHANES participant aged 18 and older. As mentioned previously, not all survey participants are eligible for linkage. For those aged 18 or older, variable VA_MATCH_STATUS on the Match Status File indicates whether the survey participant was linkage eligible and if they linked to a VA administrative record. Participants under age 18 are not included on any NCHS-VA Linked Data Files.

The Match Status File also includes NHIS and NHANES sample weights that have been adjusted to account for potential bias introduced for linkage ineligibility (i.e., ADJ_SAWT for NHIS, and ADJ_INTWT and ADJ_MECWT for NHANES). All participants who were ineligible for linkage (i.e., VA_MATCH_STATUS equal to 9) are given a linkage eligibility-adjusted weight value of zero. All linkage-eligible participants have linkage eligibility-adjusted weights greater than zero, and the sum of the adjusted weights sums to the same population totals as the full sample. Detailed reports on linkage eligibility can be found in the Match Rate Tables for NCHS-VA Linked Data Files. For more information on how the linkage eligibility-adjusted weights were created, see section 4.5.

Finally, the file includes a variable to represent linkage certainty. 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 probabilistic cut-off value to determine which pairs were links (an inferred match). For additional discussion on how PROBVALID was estimated, see Appendix I, sections 3.3 and 3.4. NCHS used a probabilistic cut-off value which aimed to minimize the total estimated counts of Type I error (false positive links) and Type II error (false negative links). However, because there are concerns that using pairs with low PROBVALID might be inappropriate for certain analyses of linked records, a PROBVALID threshold of 0.85 was established as the lowest threshold for the acceptance of links into datasets made available for external researchers.

In the NCHS-VA linkage, NCHS used a probabilistic cut-off value of 0.85 to determine final match status. Candidate pairs with a PROBVALID that exceeded the probabilistic cut-off (i.e., PROBVALID>0.85) were deemed a link. The estimated Type I error was 0.05% and the Type II error was 0.70% when applying the PROBVALID>0.85 threshold. For additional discussion on cut-off determination and record selection please see Appendix I, section 4. For some analyses, it may be desirable to reduce the Type I error. In order to do this, researchers could increase the probability cut-off value (to a value closer to 1.0). Of note, the PROBVALID cannot be decreased from 0.85 (see Appendix I). To change the NCHS link acceptance cut-off value, researchers should request the variable PROBVALID in their Research Data Center (RDC) proposal (see section 4.1).

Detailed descriptions for the complete list of variables contained in each of the NCHS-VA Linked Data Files can be found in the data dictionaries available on the NCHS Data Linkage website: https://www.cdc.gov/nchs/data-linkage/VA-restricted.htm.

3.3.2. Service Record File

The Service Record File includes information on Veteran service records, including the branch of service, character of service at separation (e.g., 'honorable, general, honorable for VA'), and the era of service, through FY 2020. This file also includes race and ethnicity information from VA administrative records.

Overall, the Service Record File contains three topic areas: branch of service and separation, date and era of service, and race and ethnicity. These topic areas and the variables included in the file are described below.

Branch of service and separation – The variables in this topic area include the branch of service at last separation, characterization of separation from service, and indicators for retirement status and type.

Branch of service is only captured for the last separation. Categories for branch of service include Army, Navy, Air Force, Marines, and Other/Unknown. Variables indicating military retirement status and type of retirement are also available; however, not all Veterans are eligible for military retirement.

The Veteran's characterization, across all separations from service, is described by four dichotomous flags with response categories of Yes/No:

- Any discharge is honorable, general, honorable for VA: CHAR_HON
- Any discharge is bad conduct, dishonorable, dishonorable for VA: CHAR_DISHON
- Any discharge is other than honorable: CHAR_OTH
- Any discharge is uncharacterized/unknown: CHAR_UNK

These variable flags should not be considered indicators of eligibility for VA benefit programs, as eligibility can depend on other factors, described in section 2.3. As noted above, the characterization of discharge flags are not identified by service period (i.e., periods being defined by activation and separation dates), and the flags are not mutually exclusive.

Dates and era of service – The variables in this topic area are provided for all NCHS-VA linked participants and include the dates of first and last activation, the first and last separation, and retirement date (for retirees). The Service Record File contains an NCHS-derived variable (TIME_SINCE_SEP_CAT) to categorize the time since the Veteran's last separation and the survey interview. The categories for this variable are up to 5 years before, 5 through 10 years before, or more than 10 years before the survey interview. There is also a category noting if the separation was after the survey interview date. This derived variable only reflects the time since the Veteran's last separation.

Due to disclosure risks, exact VA administrative date variables cannot be directly accessed by the researcher. Researchers can request a derived categorical variable based on the date variables. For example, upon request, a variable can be created indicating the timing of the retirement date in relation to the interview date.

There are also binary flags which indicate whether the Veteran was in active-duty service during war eras (e.g., active-duty service during Gulf War Era) or peacetime eras (e.g., peacetime period 1955-1964). Due to small cell sizes for Veterans serving during peacetime eras, researchers may consider combining the separate peacetime variables into a single peacetime era variable.

Race and ethnicity –The variables in this topic area include two separate variables indicating race (VA_RACE) and ethnicity (VA_HISPANIC). This information was obtained from many sources, including purchased data (which may be imputed through a commercial algorithm).[22] Therefore in the VA administrative data, the assignment of race or ethnicity may be different from the survey respondent report. It is recommended that VA_RACE and VA_HISPANIC only be used for methodological research, for example comparing their values with reported race and ethnicity variables from the NCHS surveys. For other research requiring race and ethnicity it is recommended that data users obtain this information from the NCHS survey files (see section 4.2).

3.3.3. VA Utilization File

The VA Utilization File includes information on the timing of VA benefit receipt, service-connected disability and indicators of VA benefit utilization from FY 2005 through FY 2020. The VA Utilization File contains a single record for each survey participant that linked to VA administrative data. Survey participants that were not eligible for linkage or were eligible and did not link to a VA administrative record are not included in the VA Utilization File.
Overall, the VA Utilization File contains two topic areas, VA administrative information and FY benefits. These topic areas and the variables included in the file are described below.

VA administrative information – The variables in this topic area include variables on VA health care enrollment priority rating (PRIO1_8_FYXX), the number of service-connected disabilities (NUMBER_OF_SC_CONDITION_FYXX), and total combined disability rating (TOTAL_COMBINED_RATING_FYXX). These three variables are not restricted to those enrolled in VA health care or those utilizing VA benefit programs. These variables can be populated for any Veteran who has been assessed or initiated application for benefits with the VA, and each fiscal year can have a unique value. Enrollment in VA health care for a specific fiscal year is indicated by variable IN_ENR_FYXX.

Finally, gross and net monthly compensation and pension payment amounts are available through variables GROSS_AWARD_AMOUNT_FYXX and NET_AWARD_AMOUNT_FYXX. The gross amount is the payment prior to deductions.

FY benefits – The variables in this topic area include indicators of utilization of any of the following VA benefit programs in each FY. They can be identified using the variable VA_BENEFIT_FYXX which includes health care, service-connected disability compensation, pension, VA guaranteed home loan, life insurance, education, training, and veteran readiness (vocational rehabilitation) and employment benefit programs.

[22] USVETS, Data Governance & Analytics, Office of Enterprise Integration, Department of Veterans Affairs.

Lastly, for each benefit (including VA health care) there are additional variables to indicate the type of benefit program utilization in a specific fiscal year:

- Health care (IN_VHA_FYXX)
- Service-connected disability compensation (COMP_FYXX)
- Pension (PENS_FYXX)
- VA guaranteed home loan (IN_LGY_FYXX)
- Life insurance (LIFE_INS_USER_FYXX)
- Education, training, veteran readiness (vocational rehabilitation) and employment (EDUC_IN_VRE_FYXX). Note that these benefits are grouped together into one variable.

Burial and memorial services used by service members, or their families, are not included in the summary variable VA_BENEFIT_FYXX or in the individual benefit variables included in the VA Utilization File.

4 Analytic Considerations

This section summarizes some key analytic issues for users of the NCHS-VA Linked Data Files; however, it is not an exhaustive list. This document will be updated as additional analytic issues are identified and brought to the attention of the NCHS Data Linkage Team (datalinkage@cdc.gov).

4.1 Access to the Restricted-Use NCHS-VA Linked Data 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 made available in secure facilities for approved research projects. Researchers who wish to access the NCHS-VA Linked Data Files must submit a research proposal to the NCHS RDC to obtain permission to access the restricted-use files. All researchers must submit a research proposal to determine if their projects are feasible and to gain access to these restricted data files. The proposal provides a framework which allows RDC staff to identify potential disclosure risks. More information regarding the RDC and instructions for submitting an RDC proposal are available from: https://www.cdc.gov/rdc/ (accessed September 13, 2022).

4.2 Variables to Request in RDC Proposals

To create analytic files for use in the RDC, a researcher provides a file containing the variables from the public-use NCHS survey data to the RDC for merging with the variables from the NCHS-VA Linked Data Files, and any requested restricted use variables from NCHS surveys. The exact variables from the NCHS-VA Linked Data Files, and any restricted-use variables from NCHS surveys, that the researcher intends to use also need to be specifically requested as part of a researcher’s application to RDC. Staff in the RDC verify the full list of variables (restricted and public use) and check for potential disclosure risk.

It is recommended that researchers request the following variables, available from the public-use NCHS survey files, for inclusion in analytic files:

- Sample weights and design variables—These variables are needed to account for the complex design of the NCHS surveys. The names of the weights and design variables differ depending on which NCHS survey is being used. These can be identified using the documentation for each NCHS survey and the Match Status File for the NCHS-VA Linked Data Files. As discussed in section 3.3.1 and below, NCHS recommends using the linkage eligibility-adjusted sample weights provided with the linked data to account for linkage eligibility bias.

- Demographic information—It is recommended that users who require demographic information on participants (such as sex, age, race and ethnicity) obtain this information from the NCHS survey files. The names of the demographic variables differ depending on which NCHS survey is being used and can be identified using the documentation for each NCHS survey.
- Socio-economic information—It is recommended that users who require socio-economic information on participants (such as income and education) obtain this information from the NCHS survey files. The names of the socio-economic variables differ depending on which NCHS survey is being used and can be identified using the documentation for each NCHS survey.

Although the complete list of variables used for specific analyses differs, the following variables from NCHS surveys should be considered for inclusion:

- Geography—Geography information is not available on the administrative data for linked participants. It is recommended that users who require information on geography request it from the NCHS survey. Some variables on geography (such as variables for region) are available in the NCHS public-use files, and others are available on request from NCHS survey restricted-use files.[23,24]
- Dates—Researchers who are interested in analyzing VA benefit program utilization data that is concurrent to, before, or after the survey interview (or exam for NHANES) should request the following variables:
    - NHIS calendar quarter and year of interview—The calendar quarter (INTV_QRT) and year of interview are available on the NHIS public-use files. For example, researchers may be interested in using these variables if information on timing of the survey and VA benefit program utilization are of interest. For more information on temporal alignment of VA Benefit Program Utilization see section 4.6.
    - NHANES month and year of examination and interview—The year and month of a survey participant's interview and examination can be requested from NHANES restricted files.[19] For example, researchers may be interested in using these variables if information on timing of the survey and/or exam, and VA benefit program(s) utilization is of interest, however, the exact year and month (and derived FY variables) for NHANES data cannot be removed from the RDC due to disclosure concerns. For more information on temporal alignment of VA Benefit Program Utilization see section 4.6.

4.3 Merging NCHS-VA Linked Data Files with NCHS Survey Data

To perform record-level analysis, the restricted-use NCHS-VA Linked Data Files can be used in conjunction with the NCHS survey data (described above in section 2.1 and 2.2). A unique survey

[23] NHIS Restricted Variables. https://www.cdc.gov/rdc/b1datatype/dt1225.htm (accessed September 13, 2022).
[24] NHANES Restricted Variables. https://www.cdc.gov/rdc/b1datatype/dt1222.htm (accessed September 13, 2022).

participant identification variable must be created and included for each public use survey file, to allow analysts to merge survey data for survey participants with their information from the NCHS-VA Linked Data Files. These identifiers are survey-specific and may be constructed differently across survey years. Please refer to Appendix II: Merging NCHS-VA Linked Data Files with NCHS Survey Data for guidance on identifying and constructing (if necessary) the appropriate identification variable for merging survey data and the NCHS-VA Linked Data Files.

4.4 2005-2018 NHANES Sample Sizes

The total number of linked participants from 2005-2018 NHANES is 3,523. Researchers should note that the sample size for females 65 and older does not meet the disclosure review threshold when broken out by survey cycle. Researchers should consider this limitation when requesting to analyze data by sex and age groups.

4.5 Linkage Eligibility-adjusted Participant Survey Weights

The sample weights provided in NCHS population health survey data files adjust for oversampling of specific subgroups and differential nonresponse and are post-stratified to annual population totals for specific population domains to provide nationally representative estimates. The properties of these weights for linked data files with incomplete linkage, due to ineligibility for linkage, are unknown. In addition, methods for using the survey weights for some longitudinal analyses require further research. Because this is an important and complex methodological topic, ongoing work is being done at NCHS and elsewhere to examine the use of survey weights for linked data analysis.[25] More detailed information on adjusting sample weights for linkage eligibility can be found in this NCHS report.[20, 26]

NCHS has created linkage eligibility-adjusted sample weights for the linked NCHS-VA files, available in the Match Status File (see section 3.3.1). The choice of which adjusted sample weight to use depends on the analysis and, more specifically, on the variables used in the analyses and the survey years included. Below are important considerations for the two surveys.

For NHIS: As only sample adults were included for linkage in the 2005-2018 NHIS, linkage eligibility-adjusted analyses of 2005-2018 NHIS sample adult participants should incorporate the linkage eligibility-adjusted sample adult weights (ADJ_SAWT).

For NHANES: Analyses should incorporate either the linkage eligibility-adjusted interview weights (ADJ_INTWT) or, if analytic variables are based on data obtained during the MEC examination, the linkage eligibility-adjusted MEC examination weights (ADJ_MECWT).

If researchers wish to further adjust sample weights this can be done using the VA_MATCH_STATUS variable from the Match Status File to determine linkage eligibility (see section 3.3.1).

4.6 Temporal Alignment of VA Benefit Program Utilization

Data from NCHS surveys have been linked to multiple years of VA administrative data. Depending on the survey year and the Veteran’s participation in VA benefit programs, indicators of VA utilization may be available for survey participants concurrent to, before, or after the

[25] Aram J, Zhang C, Golden C, Zelaya CE, Cox CS, Ye Y, Mirel LB. Assessing Linkage Eligibility Bias in the National Health Interview Survey. Vital Health Stat 2. 2021 Mar;(186):1-28. PMID: 33663652.

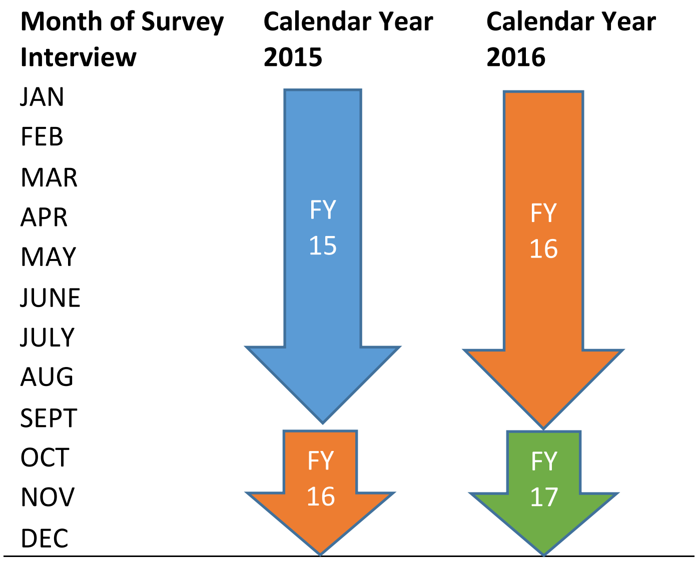
[26] Golden, C., et al., Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare Medicaid Services. Vital Health Stat 1, 2015(58): p. 1-53.

survey interview (or exam for NHANES) date. Due to disclosure risks, information on date of interview (or exam date for NHANES) cannot be directly accessed by the researcher, but upon request, RDC staff can derive categorical variables for researchers to use in the RDC.

While the surveys are conducted on a calendar year basis, VA administrative data are organized by fiscal year, which begins October 1st and ends September 30th. To assist in the calculation of alignment between the NHIS interview and the fiscal year of VA benefit program utilization, the variable SURVEY_FY, indicating the fiscal year in which the survey occurred, is provided on the VA Utilization File. This variable was created using the restricted-use variable date of interview. For example, if the NHIS interview occurred in Quarter 4 of 2015, then the corresponding VA fiscal year would be 2016. Therefore, if a data user was interested in VA benefit program utilization concurrent to survey year 2015, they would need to request and analyze the VA Utilization File variables from FY 2015 and FY 2016.

The interview and exam dates for NHANES are not released to researchers, and NHANES occurs on a two-year survey cycle defined by calendar years. Due to the misalignment of calendar and fiscal year, a researcher will need to request a three-year range of fiscal year variables in order to access benefit program utilization information concurrent to the NHANES cycle (see Figure 1 below). For example, to analyze concurrent information for a 2015-2016 NHANES cycle, a researcher should request FY 2015, FY 2016, and FY 2017 variables of interest. As noted above, due to disclosure concerns, the output from the RDC cannot indicate the exact fiscal year of occurrence for NHANES participants.

Figure 1. Relationship of Calendar Year and Fiscal Year (FY)



5 Additional Related Data Sources

Each of the NCHS surveys that have been linked to the VA administrative data have also been linked to death information obtained from a linkage with the National Death Index (NDI). The linked NDI mortality files include information on the date and cause of death for linked decedents and provide the opportunity to conduct outcome studies designed to investigate the association of a wide variety of health factors with mortality. More information about the NCHS-NDI linked mortality files can be found at: https://www.cdc.gov/nchs/data-linkage/mortality.htm.

NCHS has also previously linked NHIS and NHANES data to Center for Medicare & Medicaid Services (CMS) Medicare and Medicaid enrollment and claims data. The linked Medicare and Medicaid files provide information on program enrollment, health care utilization for covered services, as well as prescription drug data. Combining the linked VA, Medicare and Medicaid files will provide researchers with more detailed information regarding a Veteran's use of health care services that are covered by Medicare and/or Medicaid. More information regarding which NHIS and NHANES years/cycles and linked Medicare and Medicaid administrative data are available for research use in the RDC is available at https://www.cdc.gov/nchs/data-linkage/medicare.htm and https://www.cdc.gov/nchs/data-linkage/medicaid.htm.

NCHS also recently completed a linkage of NHIS and NHANES data to federal housing assistance program data obtained from the Department of Housing and Urban Development (HUD). The linked HUD administrative data files include variables pertaining to the recipient's participation in HUD's Housing Choice Voucher (HCV), Public Housing (PH), and/or Multifamily (MF) housing assistance programs. More information regarding the linked NCHS-HUD data files can be found at: https://www.cdc.gov/nchs/data-linkage/hud.htm.

Researchers may request variables from the NHIS and NHANES linked NDI, CMS Medicare and Medicaid, and HUD data files in their RDC proposals. Each of these files can be merged with the NCHS-VA Linked Data Files using the survey-specific unique participant identification variable (see Appendix II).

Appendix I: Detailed Description of Linkage Methodology

1 NCHS and VA Linkage Submission Files

A submission file is a dataset specially prepared for submission to the linkage analysis process, by having all necessary variables and records correctly formatted for this. Submission files, which contained the cleaned and validated PII fields, were separately created for NCHS survey records and for VA administrative records. To accomplish this, there were an initial series of processes that performed various data cleaning routines on the PII fields within each of the separate files containing NCHS survey and VA administrative records, prior to their linkage. Of note, processing was conducted separately for NCHS and VA records. The following PII fields were individually processed and output to its own file (i.e., there were separate files for SSN, DOB, name, etc., each record showing a possible value for that field for each survey participant or Veteran administrative record:

- SSN (validated)[27] [28]
- DOB (month, day, and year)
- Sex
- 5-Digit ZIP code and state of residence
- First, middle initial, and last name

Identifier values deemed invalid by the cleaning routine were changed to a null value. Also, each of the routines involved very basic checks related to specific characteristics of the variable to which it was applied. A few examples where this occurred include:

- Date values: when invalid or outside of expected range, they are set to null
- Sex values: when multiple sex values are seen for the same person, sex is set to null
- 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—it is common for hospitals to use the mother’s first name when no name has been decided for the baby. Name parts (i.e. first name or last name) that contain specific keywords such as baby, baby boy, baby girl, BB, BG, etc. are changed to missing.
    - Nulling of Jane/John Doe
    - Removal of titles such as Mister, Miss, etc.
    - Removal of suffixes such as Junior, II, etc.
    - Removal of special text unique to survey such as first name listed as “Void”

To increase the likelihood of finding a link, multiple or alternate submission records were used for each linkage eligible NCHS survey participant based on variations of the linkage variables. VA

[27] Complete 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 is not 012345678. For some surveys and survey years, only the last 4-digits (SSN4) were collected from survey participants.

[28] If SSN missing or invalid, then SSN was extracted from their Health Insurance Claim Number (HICN), if provided. SSN was extracted from the HICN only if the survey participant was identified as the primary claimant for Medicare benefits.

records could be matched to any or all of the submission records created for a survey participant. Similar to the cleaning process, a more elaborate routine was used to generate alternate records involving the name fields. For survey participants with multiple name parts, common nicknames, and for common Hispanic and Asian names, additional records were generated using each individual piece as a possible name value. For example, the name “Beth” may be a nickname for a formal name like “Elizabeth.” In this situation, a record for “Beth” and a record for “Elizabeth” were created and submitted for linkage. NCHS created a common nickname lookup file which was used to generate a second record replacing the nickname with the formal name. Table 1 below provides two examples of how multiple part name information was used to generate alternate records, using hypothetical data. For survey participant A, the first name was used to generate multiple records, and for survey participant B, the last name was used.

Table 1. Example of Alternate Record Generation Using Name Fields

Participant ID	First Name	Middle Initial	Last Name	Alternate Record
A	John H		Smith	0
A	John	H	Smith	1
A	H		Smith	1
A	John		Smith	1
B	John	R	Smith Jones	0
B	John	R	Smith	1
B	John	R	Jones	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 separately created for NCHS survey records and for VA administrative records. During this process, multiple submission file records were created for each participant/administrative record to show all combinations of the recorded values for these fields. That is, if a participant/administrative record 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/administrative record (see Table 2 for example). Submission records that did not meet the eligibility requirements (see section 3.1) were removed from the submission file.

Table 2. 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.

2 Deterministic Linkage Using Unique Identifiers

The deterministic linkage, which was the next step in the linkage process, used only the NCHS and VA submission records that included a valid format SSN[29]. The algorithm performed two passes on the data, first checking for full SSN9 agreement and then for records where the SSN4 agreed. 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, ZIP code of residence, and state of residence. If the ratio of agreeing identifiers to non-missing identifiers was greater than 50% (1st pass using SSN9) or greater than 2/3 (2nd pass using SSN4), 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 5 non-missing PII variables in agreement to be deemed a deterministic match. Of note, NCHS survey participants were excluded from the second pass (i.e., using the SSN4) if they were deterministically linked in the first pass. 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. To infer which pairs of records 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 data linkage expert Peter Christen, blocking or indexing, "splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key)."[30] 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 block 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' as the validation dataset and a sample of survey and administrative submission records as training

[29] If SSN missing or invalid, then SSN was extracted from their HICN, if provided. SSN was extracted from the HICN only if the survey participant was identified as the primary claimant for Medicare benefits.

[30] Christen, P. 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 (accessed September 13, 2022).

data. For more detailed information on the supervised machine learning algorithm used please refer to "Learning Blocking Schemes for Record Linkage."[31,32]

The machine learning algorithm learned 14 blocking passes to be used in the blocking scheme. Table 3 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 a blocking variable. Further, if only the ZIP code of residence was used as a blocking variable, then state of residence was excluded from the list of scoring variables as it is implied to be in agreement on all records. Likewise, if first name was used as a blocking variable, then sex was excluded from the list of scoring variables due to high correlation between the two variables.

Table 3. Blocking and Scoring Scheme used to Identify and Score Potential Links

Key Number	Blocking Key	Scoring Key
1	Last name, month of birth, day of birth, year of birth	First name, middle initial, state of residence, ZIP code of residence, sex
2	Month of birth, day of birth, year of birth, state of residence, sex	First name, middle initial, last name, ZIP code of residence
3	Last name, first name, state of residence, sex	Middle initial, month of birth, day of birth, year of birth, ZIP code of residence
4	Last name, month of birth, year of birth, state of residence, sex	First name, middle initial, day of birth, ZIP code of residence
5	First name, month of birth, year of birth, state of residence, sex	Middle initial, last name, day of birth, ZIP code of residence
6	Last name, month of birth, day of birth, state of residence, sex	First name, middle initial, year of birth, ZIP code of residence
7	First name, month of birth, day of birth, state of residence, sex	Middle initial, last name, year of birth, ZIP code of residence
8	Last name, first name, month of birth, year of birth	Middle initial, day of birth, state of residence, ZIP code of residence
9	Day of birth, year of birth, state of residence, ZIP code of residence	First name, middle initial, last name, month of birth, sex
10	Last name, first name, day of birth	Middle initial, month of birth, year of birth, state of residence, ZIP code of residence
11	First name, month of birth, day of birth, year of birth	Middle initial, last name, state of residence, ZIP code of residence

[31] Michelson, M. and Knoblock, C.A. "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 (accessed September 13, 2022).

[32] 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 (accessed September 13, 2022).

12	Last name, year of birth, state of residence, ZIP code of residence, sex	First name, middle initial, month of birth, day of birth
13	Last name, day of birth, year of birth, state of residence, sex	First name, middle initial, month of birth, ZIP code of residence
14	Month of birth, year of birth, state of residence, ZIP code of residence	First name, middle initial, last name, day of birth, sex

3.2 Score Pairs

Next, each pair 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 below), which allowed the linkage algorithm to select final links to include in the linked file. The scoring process followed the following order:

1. Calculate M- and U- probabilities (defined below)
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
- Sex
- State of Residence
- ZIP Code (conditional on state agreement)

3.2.1 Calculate M- and U- Probabilities

The **M-probability** – the probability that the values of identifiers on a pair of records agree, given that the records represent the same person (i.e., the records are a match) – was estimated separately within each individual blocking pass. M-probabilities were calculated for each of the identifiers not used in the blocking key (Table 3). Within the blocking pass, pairs with agreeing SSN were used to calculate the M-probabilities, as these are assumed to represent the same individual. For records with a SSN9, agreeing SSN was defined as 8 or more digits being the same. For records with a SSN4, we required all 4 digits to be in agreement and at least 5 PII variables agreeing, totaling more than 2/3's agreement of all non-missing PII variables. For example, if we have a record with 6 non-missing PII variables and 5 agree, this would be kept for M-probability estimation. However, if we have a record with all 8 non-missing and 5 agree, this would not be used for M-probability estimation since it does not meet the 2/3 agreement requirement (i.e., 5/8=0.625). 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 record for each NCHS survey ID and VA US-Vet ID (see Tables 4 and 5 for an example showing alternate record summarization). Table 4 is an example of how the agreement flags for each of the scoring variables in Blocking pass 3 are created. A value of 1 means the information in the variable is exactly matching, while a 0 means they are not. Table 5 then represents how the multiple submission records in table 4 are summarized into one record for each survey and 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 5 are then used to estimate the M-probabilities for each of the specific scoring variables. For example, among qualifying pairs in table 5 for blocking pass 3, 99.4% (M-probability Day Birth=0.994) agree on day of birth and 94.5% (M-probability ZIP=0.945) agreed on ZIP code of residence.

Table 4. Example of Agreement Flags Using Blocking Pass 3 as an Example

Person Identifiers		PII Agreement flags[1]				
Participant ID	VA US-Vet ID	Day of birth	Month of birth	Year of birth	ZIP Code	Middle Initial
1	1	1	0	1	0	.
1	1	.	1	1	0	0
1	1	1	0	1	0	0
2	2	1	0	1	0	0
3	789	1	1	.	0	1
3	789	0	1	0	1	1
3	789	.	1	0	1	.
3	789	0	0	1	1	1
3	322	1	0	1	1	1

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

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

Table 5. Example Showing Summarization of Blocked Records for M-Probability Estimation, Based on Records in Table 4

Person Identifiers		PII Agreement flags[1]				
Participant ID	VA US-Vet ID	Day of birth	Month of birth	Year of birth	ZIP Code	Middle Initial
1	1	1	1	1	0	0
2	2	1	0	1	0	0
3	789	1	1	1	1	1
3	322	1	0	1	1	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, . = missing values

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

- First/last initial agreement – used in the scoring process when only an initial was present in the name field

- Jaro-Winkler Similarity Levels – this process is explained in greater detail in section 3.2.2
- ZIP Code of residence – because ZIP codes are dependent on the state in which they are located, only pairs of records where state of residence agreed were used in the computation of the ZIP code M-probability (i.e., if state was not in agreement then it would be assumed that ZIP code would also not agree).

The **U-probability** - 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 only calculated 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 SSN were not in agreement (defined as having less than 5 matching digits for records with SSN9 values and if any digits was not in agreement for records with SSN4 values). 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 that had majority of the non-missing PII among scoring variables were in agreement) were excluded prior to calculating the U-probabilities. For example, when computing the U-probability for day of birth in blocking pass 12, records that did not agree on SSN that had majority of the PII among first name, middle initial, and month of birth were excluded from the assumed non-matches. These records were assumed to be probable matches given that a majority of the PII between the survey and administrative records were in agreement.

The U-probabilities, however, were calculated for each value (level) of a variable. For example, the state of residence U-probabilities within blocking pass 1 for Florida and Pennsylvania were, 0.052 (5.2%) and 0.091 (9.1%), respectively. However, for first and last name, the U-probabilities were calculated in a different manner further described in section 3.2.2.

3.2.2 M- and U-Probabilities for First and Last Names

Similar to the M-probability, Jaro-Winkler levels (0.85, 0.90, 0.95, and 1.00) were calculated for use in the U-probability computation. Because agreement levels fall over a range, first and last name U-probabilities were computed for each Jaro-Winkler score level. Since there are a plethora of possible values for first and last name (i.e., one for each possible name), it was impractical to compute U- probabilities for a specific name 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 submission file and a simple random sample of 5% of records with non-missing name information of the VA submission file.

Complete name tallies (separately, for first and last names) were then produced for the NCHS survey submission file. For each level of name on the file, 100,000 names were randomly selected from the VA submission file 5% sample to compare to it. 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. The number of names in agreeance of the 100,000 randomly selected VA file names that agreed at that level for each name were then tallied.[33,34,35]

3.2.3 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:

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

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

Implied by the name, agreement weights were only assigned to the identifiers that have agreeing values. Similarly, non-agreement weights were only assigned to identifiers that have non-agreeing values. A non-agreement weight was always a negative value and reduced the pair weight score.

3.2.4 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 follow the same general process:

- Start with a pair weight of 0.
- Identifier agrees: add identifier-specific agreement weight into pair weight
- Identifier disagrees: add identifier-specific non-agreement weight (which has a negative value) into pair weight
- Identifiers cannot be compared because one or both identifiers from the respective records compared were missing: 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 scores 0.85 and below a disagreement weight. The algorithm assigned all 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.

[33] 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.

[34] 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.

[35] Resnick, D., Mirel, L.B., 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 September 13, 2022).

3.3 Probability Modeling

A probability model, developed from a partial expectation-maximization (EM) 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 represented the probability 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 survey participant's IDs that have linked to multiple administrative records
- Select final matches based on a probability threshold (discussed in the following section 4)

The partial EM 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.

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

   Note that in the first iteration, it was assumed that $N_{\widehat{matches,B}}$ = $N_{\widehat{non-matches,B}}$, resulting in $Adj_B = 0$. If, however, in a later iteration, the number of matches was estimated to be, $N_{\widehat{matches,B}}$ = 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.

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

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

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

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 $N_{\widehat{matches},B}$ 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.[36]

To remedy this, before the algorithm adjudicated the matches against the probability threshold, one final adjustment was made to the match probabilities (for probabilistic pairs). For pairs that had an SSN on both the NCHS survey and VA administrative record, the estimated probability was adjusted based on the last four digits of the SSN.[37]

[36] The M-probability for the last 4-digits of SSN is estimated as the rate of SSN agreement for records with high estimated match probabilities, where SSN agreement is defined as having all 4-digits in agreement between the NCHS survey and VA administrative record. The U-probabilities are estimated as the random chance that a 4-digit SSN value will agree, or simply $\frac{1}{9{,}999} \approx 0.0001$.

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

When the last four digits of SSN[38] 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 or VA administrative record. So, for these pairs:

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

4 Estimate Linkage Error, Set Probability Threshold, and Select Matches

4.1 Estimating Linkage Error to Determine Probability Cutoff

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, the percentage of them who were not true matches
- Type II Error: Among true matches, the percentage who were not linked

Because all records were included in the probabilistic linkage (i.e., even deterministic links), SSN agreement status (defined as 7 or more matching digits for records with SSN9 values and all 4 digits for records with SSN4 values) 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 SSN available on both the survey and administrative 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. Since a sizeable proportion of links were derived from the deterministic method, this had the effect of reducing the estimated Type I error by the proportion of probabilistically determined linkages among all linkages. For this linkage, the Type

[38] Rather than using the entire SSN, the last four digits are used since the first five digits of an SSN are not truly random. Prior to 06/25/2011 the first three digits represented the state where the Social Security Administration (SSA) paperwork was submitted to obtain an SSN. The fourth and fifth digit are known as a group number that cycles from 01 to 99. This additional pair weight allows for more accurate adjudication of links where other PII may not provide a clear indication of match status.

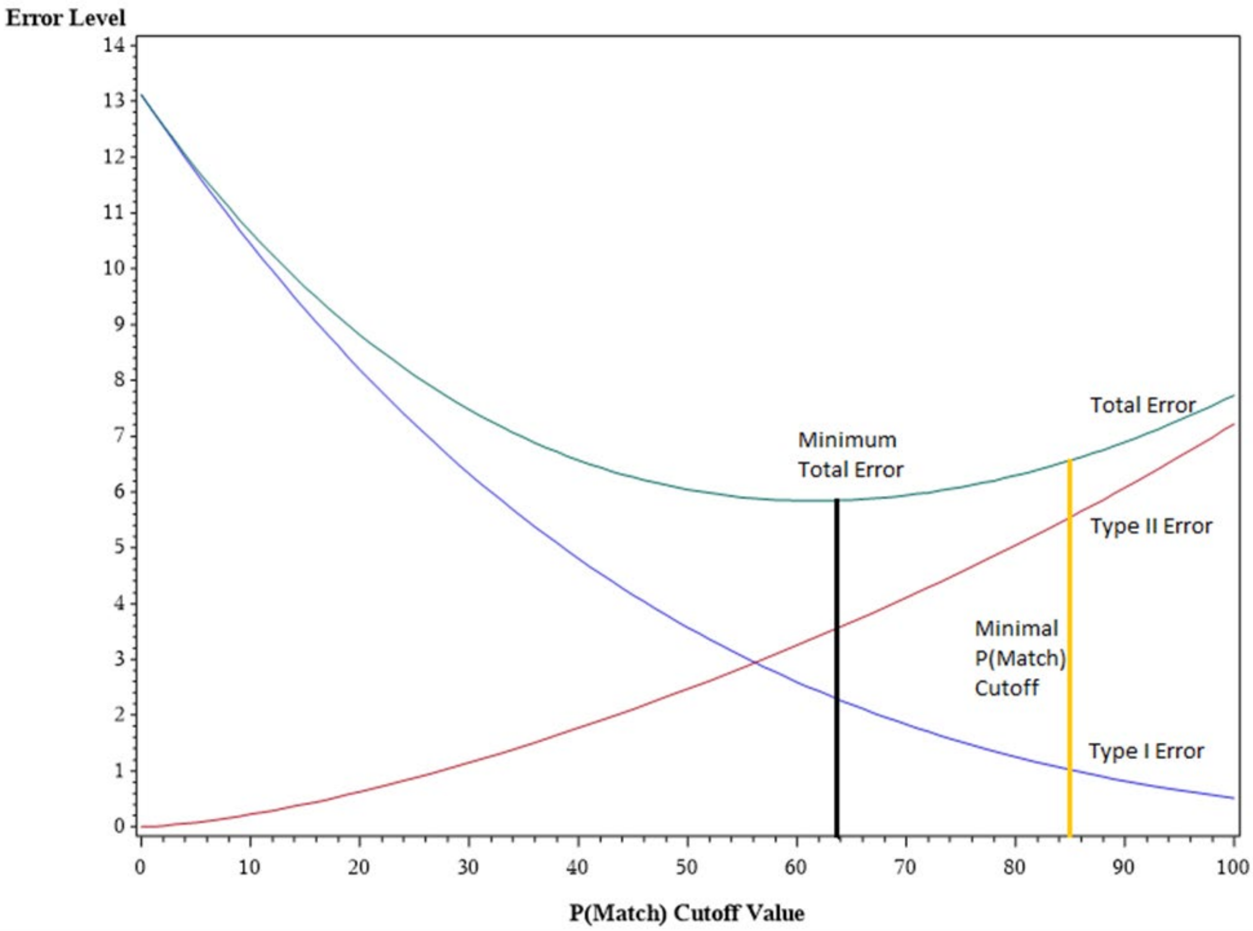
I error rate was estimated for probabilistic links as 0.16% and 32% of all links were derived from probabilistic analysis, resulting in an estimated Type I error rate for the combined linkage process of (0.32*0.0016) = 0.0005 or 0.05%.

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 9-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). 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. Similar to Type I error, adjustment was made to this error based on the fact that links having agreeing SSNs were to be linked deterministically even if they are not returned by the probabilistic approach. For example, say that the probabilistic approach was able to return 97% of true matches as links, but 50% of true matches cannot be deterministically linked (i.e., because they do not have two SSN values to facilitate a join). Then, only half of the true matches were susceptible to linkage error and the estimated Type II error rate is ½ of (1 – 0.97) = 0.015 or 1.5%. Again, as with the estimation of Type I error, it was assumed that the rate of non-linkage was identical for all records and those in the truth source. This may have been unrealistic as it might have been expected that truth source records were more readily linkable (probabilistically, but in the absence of having two SSNs) compared to all candidate pairs in general.

4.2 Set Probability Cutoff

One goal of record linkage is to have the lowest errors possible. However, as more pairs were accepted, pairs that were less certain to be matches as links increase the Type I error and decrease Type II error (see Figure 2). And as fewer pairs were accepted, pairs that were more certain to be matches as links decrease the Type I error and increase Type II error. The optimal trade-off is between Type I error and Type II error was not known, and likely this depends on the type of analysis to be conducted with the linked data, but it is assumed that it is not far from optimality 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 points and the one that showed the lowest estimate of total error was selected. For this linkage, the probability cutoff was set to 0.85. Although 0.85 did not minimize the total error, it was chosen because there are concerns that using pairs with low PROBVALID might be inappropriate for certain analyses of linked records. Therefore, PROBVALID = 0.85 was established as the lowest threshold that will be used for the acceptance of links into datasets made available for external researchers.

Figure 2: Error Level by Cutoff Value
(Schematic: not based on actual analysis)



4.3 Select Links Using Probability Threshold

The final step in the linkage algorithm was to determine links, which were pairs imputed to be matches. Links were pairs where the $Probvalid_{SSN_{Adj}}$ exceeded the set probability threshold (from section 3.2). All pairs with an adjusted probability that fell below the set probability threshold were not linked.

Following link determination, the algorithm selected the best link for each NCHS survey participant (if more than one link existed). The algorithm carried out this process by selecting the link with the higher match probability. In the event that there was a tie for the top match probability, the algorithm selected the link with the best matching SSN. If a tie still remained, the algorithm then randomly selected one of the links.

4.4 Computed Error Rates of Selected Links

Overall, the Type I and Type II linkage error rates for the NCHS-VA Linked Data Files were 0.05% and 0.70%, respectively.

Appendix II: Merging NCHS-VA Linked Data Files with NCHS Survey Data

The restricted-use NCHS-VA Linked Data Files are merged with the public-use NCHS survey data files using unique person identifiers. Therefore, it is important for researchers to include the correct survey person identification number: PUBLICID (for NHIS), or SEQN (for NHANES) in their RDC proposal (see section 4.1). For using NHIS data, it also is important to note in the descriptions below that the variable names and locations needed to construct PUBLICID vary by NHIS year.

Note: 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 a survey-specific public identification number (PUBLICID or SEQN) on any analytic files sent to the RDC. The RDC will merge data (using PUBLICID or SEQN) from the linked VA files to the analyst's file. The merged file will be at the RDC and made available for analysis.

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

NHIS 2005 – 2018

Taken together, the data items 'Survey year' (SRVY_YR), 'Household number' (HHX), 'Family number' (FMX), and 'Person number' (FPX) identify a participant within each NHIS. These data items must be concatenated to obtain the unique personal identifier (PUBLICID) used in the NHIS-VA Linked Data Files.

Variable	Location	Length	Description
SRVY_YR	3-6	4	Year of interview
HHX	7-12	6	Household serial number
FMX	16-17	2	Family number
FPX	18-19	2	Person number

SAS example:

```
length publicid $14;

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

Stata example: (note this will convert the variables to a string variable)

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

NHANES 2005-2018

Variable	Length	Description
SEQN	8	Participant identification number

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