The Linkage of the 2016 National Hospital Care Survey Data to United States Department of Veterans Affairs Administrative Data:

Linkage Methodology and Analytic Considerations

Data Release Date: December 9, 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 2016 National Hospital Care 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: www.cdc.gov/nchs/data-linkage/nhcs-va.htm

Table of Contents

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

NHCS, National Hospital Care 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 and healthcare establishment surveys, including the National Hospital Care Survey (NHCS), https://www.cdc.gov/nchs/nhcs/index.htm (accessed December 9, 2022).

The NHCS collects electronic health records or health care claims data from participating hospitals drawn from a national sample frame of non-institutional and non-federal hospitals with six or more staffed inpatient beds. Participating hospitals are requested to send all patient ambulatory care and inpatient (IP) encounters occurring within the data collection calendar year. The NHCS includes detailed information about patient's characteristics, conditions, and treatment at each participating hospital. Even though NHCS is an establishment survey (i.e., hospitals are the sampling unit) it collects patient personally identifiable information (PII), which enable data linkages.

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 the NHCS by augmenting it with administrative data collected by VA. **This report will describe the linkage of data from the 2016 NHCS to VA administrative data through September 30, 2020 (fiscal year 2020).** This linkage, collectively referred to as the NHCS-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 among patients seen at participating hospitals.

This document describes the first linkage conducted between the NHCS 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 December 9, 2022).
[2] NCHS. Restricted-Use NCHS-VA Data. https://www.cdc.gov/nchs/data-linkage/va-restricted.htm (accessed December 9, 2022).

2 Data Sources

2.1 National Hospital Care Survey (NHCS)

The NHCS is an establishment survey that collects IP, emergency department (ED), and outpatient department (OPD) episode-level data from sampled hospitals. NHCS is one of the NCHS National Healthcare Surveys, a family of surveys that are provider-based, covering a broad spectrum of health care settings (https://www.cdc.gov/nchs/dhcs/index.htm, accessed December 9, 2022).

The goal of NHCS is to provide reliable and timely healthcare utilization data for hospital-based settings, including prevalence of conditions, health status of patients, and health services utilization. NHCS collects data from participating hospitals on all IP and ambulatory care visits occurring during the calendar year for patients of all ages (including newborns). During the 2016 data collection, hospitals were given the option of providing their data in the form of electronic health records (EHRs) or as Uniform Bill (UB)-04 administrative claims records. Thus, participating hospitals provided data in the form of UB-04 claim records or EHR data, where the EHR data in 2016 were provided in the form of Continuity of Care Documents (CCDs) or custom extracts. NHCS collects patient PII (e.g., full name, date of birth, and Social Security Number (SSN)), which allows for the linkage of episodes of care across hospital units as well as to other data sources, such as VA data. The linkage described throughout this document includes only the linkage to VA data for patients with either IP or ED visits reported in NHCS; patients who only had other, non-ED OPD visits reported in NHCS have been excluded from the linkage.

The NHCS sample frame includes 6,622 non-institutional, non-federal hospitals with six or more staffed inpatient beds. A base sample of 500 hospitals and a reserve sample of 500 additional hospitals was drawn from this frame.

In 2013, to provide estimates for ED visits with incidents of substance abuse, 81 hospitals with 500 or more staffed inpatient beds were added to the NHCS sample from the reserve sample. Thus, the hospital sample frame for the 2016 NHCS data collection was 581 hospitals.

In 2016, 158 out of the 581 sampled hospitals provided data and of the 158 participating hospitals, 142 were determined to be in-scope for linkage. Hospitals were determined to be out-of-scope for linkage if they did not provide patient PII, provided less than 50 patient encounter records or did not provide patient records covering at least 6 months of the data collection period. Of those 142 linkage eligible hospitals, 140 hospitals submitted IP data and 121 hospitals submitted ED data.

2.2 VA Benefit Programs and Data

The VA provides lifelong benefits to eligible military Veterans and their families. 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.[3, 4] 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

[3] VA Health Care. https://www.va.gov/health-care/ (accessed December 9, 2022).
[4] VHA. https://www.va.gov/health/ (accessed December 9, 2022).

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[5].

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[6], pension[7], VA guaranteed home loans[8], life insurance[9], education and training[10], veteran readiness (vocational rehabilitation)[11], and other benefits.[12]

2.2.1 VA Administrative Data

VA administrative data contained in this linkage include 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 NHCS-VA Linked Data Files.

The VA administrative data included in the NHCS-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.[13] USVETS provides a comprehensive picture of the Veteran population to support statistical, trend, and longitudinal analysis. Not all information is sourced directly from VA 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 whose active-duty service was prior to 1970, may not be complete. This linkage includes VA administrative data through fiscal year (FY) 2020.

[5] VA. Eligibility for VA health care. https://www.va.gov/health-care/eligibility/ (accessed December 9, 2022).
[6] VA. Compensation. https://www.benefits.va.gov/compensation/index.asp (accessed December 9, 2022).
[7] VA. Pension. https://www.benefits.va.gov/pension/index.asp (accessed December 9, 2022).
[8] VA. VA Home Loans. https://www.benefits.va.gov/homeloans/index.asp (accessed December 9, 2022).
[9] VA. Life Insurance. https://www.benefits.va.gov/insurance/index.asp (accessed December 9, 2022).
[10] VA. VA education and training benefits. https://www.va.gov/education/ (accessed December 9, 2022).
[11] VA. Veteran Readiness and Employment (VR&E). https://www.benefits.va.gov/vocrehab/index.asp (accessed December 9, 2022).
[12] VA. Summary of VA Benefits. https://benefits.va.gov/BENEFITS/benefits-summary/SummaryofVABenefitsFlyer.pdf (accessed December 9, 2022)
[13] USVETS, Data Governance & Analytics, Office of Enterprise Integration, Department of Veterans Affairs.

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.[14]

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).[15] The data linkage work was performed at NCHS.

Linkage was attempted only for 2016 NHCS patients, aged 18 or older[16], that had patient records with at least two of the following three identifiers present: valid SSN[17,18], valid date of birth (month, day, and year)[19], or valid name (first, middle, and last)[20]. For example, if the PII on the NHCS patient record had no SSN, a full name, and only the year of birth, the record would be considered ineligible for linkage, as only one of the criteria (i.e., that for name) was met.

Eligibility for linkage can be identified using the variable (ELIGSTAT) on the NHCS-VA match status file. ELIGSTAT values include 0 (not eligible for linkage), 1 (eligible for linkage), and 2 (not eligible for VA linkage, based on sufficient PII for linkage but under the age of 18 (as of January 01, 2016) or missing month, day, and year of birth). Table 1 presents the total number of 2016 NHCS patients by age group and sex, the number who were eligible for linkage, the number who linked to VA administrative data, and the percentage of total sample and eligible for linkage who linked to VA administrative data.

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.[21]

---

[14] VA. National Center for Veterans Analysis and Statistics. https://www.va.gov/vetdata/ (accessed December 9, 2022).

[15] 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.

[16] Age computed in relation to January 01, 2016. In the event a patient had more than one unique date of birth present on the 2016 NHCS, the earliest date of birth was used to compute the patient's age.

[17] 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 four digits cannot be 0's (i.e., xxx-00-xxxx or xxx-xx-0000), and cannot be consecutive (i.e., 012345678 or 876543210).

[18] SSN was extracted from the patient's Health Insurance Claim Number (HICN), if provided. SSN was extracted from the HICN only if the patient was identified as the primary claimant for Medicare benefits.

[19] A date of birth is considered valid if at least two of the three date parts are valid date values.

[20] A name is considered valid if: either first or last name has two or more characters, and two of the three name parts (first, middle, and last) are non-missing.

[21] VA. VA Benefits and Health Care. https://www.va.gov/ (accessed December 9, 2022).

3.2 Overview of Linkage

This section outlines steps that were used to link the 2016 NHCS patient 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 NHCS patients 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 NHCS patient records 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.[22] 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 NHCS patient record that was deemed a match, VA extracted information from the USVETS database and sent the data to NCHS through a secure data transfer system. Table 1 highlights the linkage results.

Table 1. Linked 2016 NHCS - VA Administrative Records: Sample Sizes and Percent Linked, by Age and Sex

	Sample Size			Percent Linked	
	Total Sample	Eligible for Linkage[1]	Linked to VA Administrative Data[2]	Total Sample[3]	Eligible Sample[4]
Age[5]					
0-17	1,078,180	0	0	0	0
18-39	1,263,364	1,185,531	38,729	3.1	3.3
40-64	1,130,334	1,062,324	85,663	7.6	8.1
65 and over	740,281	696,172	108,610	14.7	15.6
Total	4,212,159	2,944,027	233,002	5.5	7.9
Sex[6]					
Male	2,597,453	1,214,473	200,528	7.7	16.5

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

Female	3,157,461	1,691,840	28,979	0.9	1.7
Total	5,754,914	2,906,313	229,507	4.0	7.9

NOTES: Data are presented at patient level.

[1]Eligibility for linkage is based upon patients age (must be 18 or older) and having sufficient PII in at least two of three data element groups: SSN, name, and date of birth. 1,642,060 patients in the 2016 NHCS did not have sufficient PII for linkage and were considered ineligible for linkage.
[2]This group includes linkage-eligible patients who linked to VA enrollment database.
[3]This percentage is calculated by dividing the number of linked patients by the number of patients in the total sample.
[4]This percentage is calculated by dividing the number of linked patients by the total number of linkage-eligible patients
[5]Age is as of January 01, 2016. Age is calculated by subtracting patient date of birth (DOB) from January 01, 2016. When more than one unique DOB was present, the earliest of the non-missing DOB was selected to compute age.
[6]Sex could not be determined for 68,251 in the 2016 NHCS due to missing data.

3.3 Description of 2016 NHCS-VA Linked Data Files

The NHCS-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.

3.3.1 Match Status File

The Match Status File can be used to identify which NHCS patients were eligible for linkage and linked to VA administrative records. The Match Status File contains a single record for each patient in the 2016 NHCS. As mentioned previously, not all NHCS patients are eligible for linkage. Researchers should use the variable ELIGSTAT when identifying patients who were eligible for linkage.

For those aged 18 or older, variable VA_MATCH_STATUS on the Match Status File indicates whether the patient was linkage eligible and if they linked to a VA administrative record. Patients under age 18 are considered not eligible for VA linkage and will have a VA_MATCH_STATUS equal to 9. NHCS patients under age 18 are only included on the Match Status File.

The file also 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 2016 NHCS-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 0.1% and the Type II error was 1.8% 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. To do this, researchers should 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 2016 NHCS-VA Linked Data Files can be found in the data dictionaries available on the NCHS Data Linkage website: www.cdc.gov/nchs/data-linkage/nhcs-va.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. The Service Record File contains a single record for each 2016 NHCS patient that linked to a VA administrative record. Patients that were not eligible for linkage or were eligible and did not link to a VA administrative record are not included on the Service Record File.

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 2.2. 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 NHCS-VA linked patients and include the dates of first and last activation, the first and last separation, and retirement date (for retirees). Researchers may use these VA service dates, along with patient

encounter dates, to create categorical variables to indicate whether the VA service event of interest occurred prior to, during, or after the patient encounter. (Note: Exact date information may not be removed from the NCHS RDC).

There are also binary flag variables 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).

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 multiple sources, including purchased data (which may be imputed through a commercial algorithm).[23] Therefore in the VA administrative data, the assignment of race or ethnicity may be different from the race reported in a patient hospital encounter record. Although patient race is reported in the NHCS data, the percent of patients with a survey reported valid race code is low. Researchers may wish to consider utilizing the race and ethnicity data present in the linked VA administrative records.

3.3.3. VA Utilization File

The VA Utilization File includes information on VA benefit enrollment, service-connected disability, and indicators of VA benefit utilization for the FY 2015-FY 2018 time period. The FY 2015-2018 variables provide researchers with information on VA benefit utilization for the period one year prior and one year after NHCS patient encounters occurring during the 2016 calendar year (note: data include FY 2018 data because of the misalignment of calendar and fiscal years, see section 4.3). The VA Utilization File contains a single record for each 2016 NHCS patient that linked to VA administrative data. Patients 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 by fiscal year – 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 by fiscal year 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

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

VA_BENEFIT_USER_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.

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, NCHS grouped these benefits into one variable.

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

4 Analytic Considerations

This section summarizes some analytic issues for users of the NHCS-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 NHCS-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 NHCS-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 December 9, 2022).

4.2 Merging 2016 NHCS-VA Linked Data Files with 2016 NHCS Data

To perform encounter-level analysis, the restricted-use 2016 NHCS-VA Linked Data Files can be used in conjunction with the 2016 NHCS data (described above in section 2.1). The unique NHCS patient identifier (PATIENT_ID) must be included in the requested variable list to allow analysts to merge survey encounter data for patients with their information from the NHCS-VA Linked Data Files.

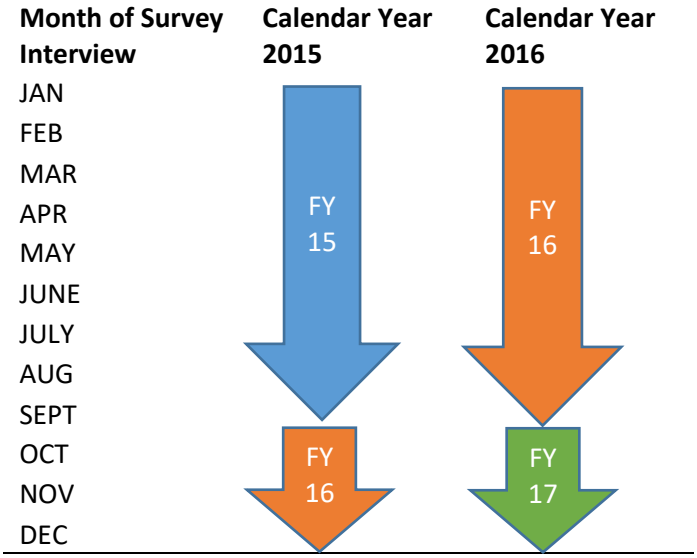
4.3 Temporal Alignment of VA Benefit Program Utilization

Data from the 2016 NHCS have been linked to multiple years of VA administrative data. The VA Utilization file contains indicators of VA benefit utilization for FY 2015 through 2018. This file

provides researchers with information about VA benefit utilization data for the FYs corresponding to one year prior and one year after any reported patient hospital encounter occurring during the 2016 calendar year.

While the hospital encounters in the NHCS occur during the calendar year (January 1, 2016, through December 31, 2016), VA benefit program utilization variables are organized by fiscal year, which begins October 1st of a given calendar year and ends September 30th of the following calendar year. Figure 1 provides an illustrative example depicting the relationship between the calendar and fiscal years.

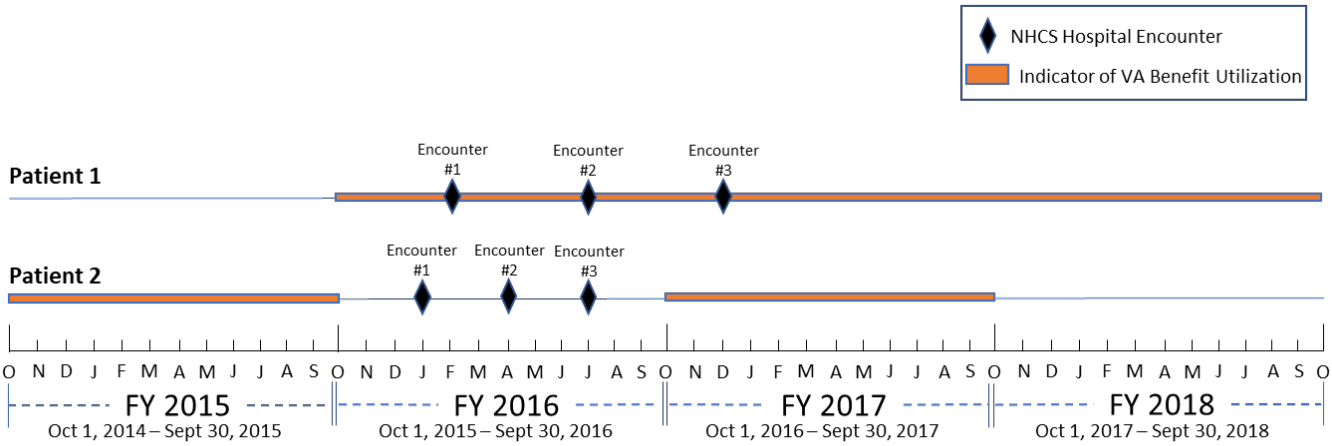
Figure 1. Relationship of Calendar Year and Fiscal Year (FY)



Since it is possible for NHCS patients to have more than one hospital encounter during the survey calendar year, the correct alignment of FY VA benefit utilization will vary by month of hospital encounter. Depending on the timing of a patient's hospital encounter and the Veteran's participation in VA benefit programs, VA utilization data may be available for NHCS patients concurrent to, before, or after their hospital encounters. Figure 2 provides an illustration of VA FY benefit period and patient encounter date alignment.

For Patient 1, researchers interested in aligning VA benefit utilization concurrent with Encounter #1 (February 2016) would utilize VA FY 2016 variables. Aligning VA benefit utilization concurrent with Patient 1 Encounter #3 (December 2016) would require the use of VA FY 2017 variables. The example for Patient 2 demonstrates the correct alignment of FY VA benefit utilization data for the time period one year prior to or after 2016 NHCS patient encounters. For Patient 2, the correct alignment for VA FY benefit utilization during the FY one year prior to Encounter #1 (January 2016) is FY2015 and during the FY after Encounter #3 (July 2016) is FY 2017.

Figure 2. Temporal alignment of 2016 NHCS encounter data and VA benefit utilization FY variables



Notes: Federal FY spans from October 1 through September 30. VA is the United States Department of Veteran Affairs. VA benefit utilization defined as having any utilization in one of the following benefit programs: health care, disability compensation, pension, VA guaranteed home loan, life insurance, education, training, and Veteran readiness (vocation rehabilitation) and employment benefit programs. For each benefit (including VA health care), there are additional variables to indicate utilization in a specific fiscal year.
Sources: NCHS, 2016 NHCS linked to VA administrative data.

5 Additional Related Data Sources

The 2016 NHCS has 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 2016 NCHCS-NDI linked mortality files can be found at: https://www.cdc.gov/nchs/data-linkage/nhcs-ndi.htm (accessed December 9, 2022).

NCHS has also previously linked 2016 NHCS data to Centers 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 linked Medicare and Medicaid administrative data are available for research use in the RDC is available at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed December 9, 2022) and https://www.cdc.gov/nchs/data-linkage/nhcs-medicaid.htm (accessed December 9, 2022).

NCHS also recently completed a linkage of the 2016 NHCS 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 2016 NHCS-HUD data files can be found at: https://www.cdc.gov/nchs/data-linkage/nhcs-hud.htm (accessed December 9, 2022).

Researchers may request variables from the 2016 NHCS linked NDI, CMS Medicare and Medicaid, and HUD data files in their RDC proposals. Each of these files can be merged with the 2016 NHCS-VA Linked Data Files using the NHCS patient identifier variable (PATIENT_ID).

Appendix I: Detailed Description of Linkage Methodology

1 2016 NHCS 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 NHCS 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 NHCS and VA administrative records, prior to their linkage. The following PII fields were individually processed and output to their own file (i.e., there were separate files created for SSN, DOB, name, etc., each record showing a possible value for that field for each survey participant or Veteran administrative record:

- SSN (validated)[24,25]
- 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 missing
- Sex values: when multiple sex values are seen for the same person, sex is set to missing
- Name values: multiple edits are applied:
  - Removal of special characters such as [“-.,<>/?, etc.]
  - Removal of descriptive words such as twin, brother, daughter, etc.
  - 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.
  - Jane/John Doe have full name (i.e., first, middle, and last) changed to missing
  - 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 NHCS patient based on variations of the linkage variables. VA records could be matched to any or all the submission records created for a NHCS patient. Similar to the cleaning process, a more elaborate routine was used to generate alternate records involving the

[24] 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 four digits cannot be 0’s (i.e., xxx-00-xxxx or xxx-xx-0000), and is not 012345678.

[25] SSN was extracted from the patient’s Health Insurance Claim Number (HICN), if provided. SSN was extracted from the HICN only if the patient was identified as the primary claimant for Medicare benefits.

name fields. For NHCS patients 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 2 below provides two examples of how multiple part name information was used to generate alternate records, using hypothetical data. For patient A, the first name was used to generate multiple records, and for patient B, the last name was used.

Table 2. Example of Alternate Record Generation Using Name Fields

Patient 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 NHCS records and for VA administrative records. During this process, multiple submission file records were created for each patient/administrative record to show all combinations of the recorded values for these fields. That is, if a patient/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 patient/administrative record (see Table 3 for example). Submission records that did not meet the eligibility requirements (see section 3.1) were removed from the submission file.

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

Patient 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 first step in the linkage process, used only the NHCS and VA submission records that included a valid format SSN[26]. The algorithm performed two

[26] SSN was extracted from the patient's HICN, if provided. SSN was extracted from the HICN only if the patient was identified as the primary claimant for Medicare benefits.

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 9-digit SSN matched. 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 all 9 digits of SSN) or greater than 2/3 (2nd pass using last four digits 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 5 non-missing PII variables in agreement to be deemed a deterministic match. Of note, NHCS patients were excluded from the second pass (i.e., using the last four digits of SSN) 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 for all records. 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)."[27] 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 data. For more detailed information on the supervised machine learning algorithm used please refer to "Learning Blocking Schemes for Record Linkage."[28,29]

---

[27] 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 December 9, 2022).

[28] 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 December 9, 2022).

[29] 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 December 9, 2022).

The machine learning algorithm learned 14 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 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 4. 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
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 in the blocks 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 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. 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 NHCS patient ID and VA US-Vet ID (see Tables 5 and 6 for an example showing alternate record summarization). Table 5 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 6 then represents how the multiple submission records in Table 5 are summarized into one record for each patient 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 6 are then used to estimate the M-probabilities for each of the specific scoring variables. For example, among qualifying pairs in Table 6 for blocking pass 3, 99.4% (M-probability Day Birth=0.994) agreed on day of birth and 94.5% (M-probability ZIP=0.945) agreed on ZIP code of residence.

Table 5. Example of Agreement Flags Using Blocking Pass 3 as an Example

Person Identifiers		PII Agreement flags[1]				
Patient 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 6. Example Showing Summarization of Blocked Records for M-Probability Estimation, Based on Records in Table 5

Person Identifiers		PII Agreement flags[1]				
Patient 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 (of 9) matching digits). 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 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 and had

majority of the PII among first name, middle initial, and month of birth in agreement, were excluded from the assumed non-matches. These records were assumed to be probable links 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 were 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 NHCS 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 NHCS 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.[30,31,32]

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:

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

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

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.

---

[30] 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.

[31] 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.

[32] 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 December 9, 2022).

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 similarity 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 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 (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 NHCS patient 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, $\widehat{N_{matches,B}}$, 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 $Odds_{1,B} = 2^{(8.4+\,-5.61)} \approx 6.9$

   if for Pair 2 of blocking pass B, the pair-weight is -2.5, then $Odds_{2,B} = 2^{(-2.5+\,-5.61)} \approx 0.0036$

   …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.[33]

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 NHCS and VA administrative record, the estimated probability was adjusted based on the last four digits of the SSN.[34]

When the last four digits of SSN[35] 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:

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

[33] The M-probability for the last four 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 four digits in agreement between the NHCS and VA administrative record. The U-probabilities are estimated as the random chance that a four-digit SSN value will agree, or simply $\frac{1}{9{,}999} \approx 0.0001$.

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

[35] 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.

No adjustment was made for pairs that did not have an SSN on either the NHCS 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

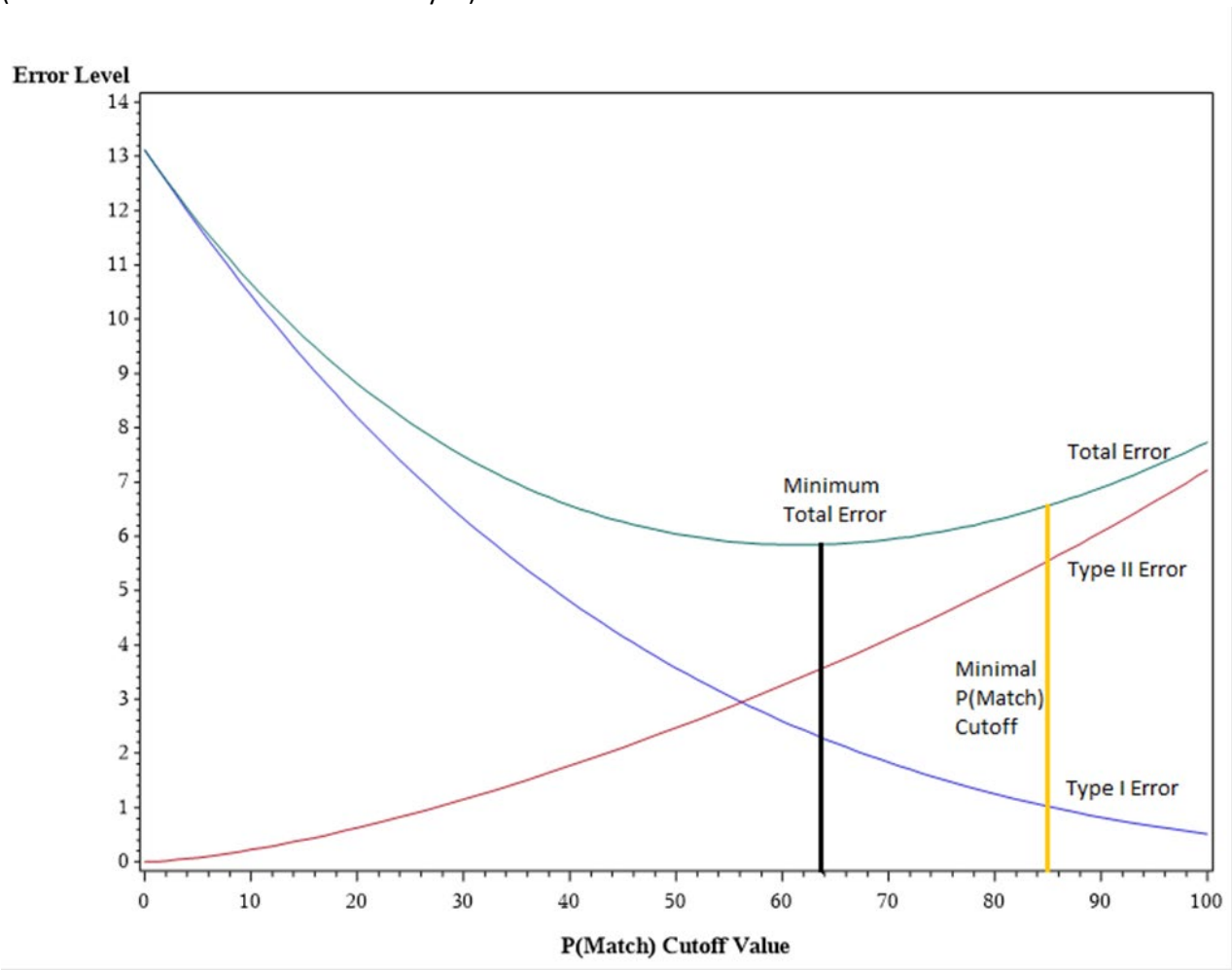
Because all records were included in the probabilistic linkage (i.e., even deterministic links), SSN agreement status (defined as 7 or more matching digits) was used to measure Type I error. Type I error for probabilistic links was measured as the total number of probabilistic links with non-agreeing 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 49% 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 I error rate was estimated for probabilistic links as 0.2% and 51% of all links were derived from probabilistic analysis, resulting in an estimated Type I error rate for the combined linkage process of (0.51*0.002) = 0.001 or 0.1%.

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). 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. 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 linkage 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 3). 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 3: 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 at or below the set probability threshold were not linked.

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

4.4 Computed Error Rates of Selected Links

Final error rates were computed for selected links (described in section 4.3). Table 7 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 2016 NHCS-VA linkages. 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 NHCS record was a match to the VA 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 that 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 7. Algorithm Results for Total Selected Links

	Cutoff	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
2016 NHCS	0.85	233,002	114,232 (49%)	118,770 (51%)	0.1%	1.8%

Table 8 provides the total selected links, number of probabilistic and deterministic links, and the estimated Type I and II error rates for the selected links, by record type source for the 2016 NHCS. As shown in Table 8, UB-04 claims have slightly higher estimated linkage error (both Type I and II) compared to the EHR records. Due to elevated levels of missing data in EHRs compared to the UB-04 claims records, the number of deterministic matches made by the algorithm for EHR Custom Extract (86.3%) is proportionally higher than UB-04 deterministic matches (50.9%). This resulted in a lower proportion of EHRs having VA data extracted based on the probabilistic linkage. Additionally, CCD data were delivered without SSN information. This resulted in 100% of CCDs having VA data extracted based on the probabilistic linkage and therefore the Type II linkage error was not calculated.

Table 8. Algorithm Results for Total Selected Links by 2016 NHCS Data Source

Data Source	Cutoff	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
UB-04 Claims	0.85	182,471	92,796 (50.9%)	89,675 (49.1%)	0.1%	1.7%
EHR Custom Extract	0.85	24,837	21,436 (86.3%)	3,401 (13.7)	<0.1%	0.5%
CCD	0.85	25,694	0 (0%)	25,694 (100%)	0.4%	*

*Unable to estimate Type II linkage error due to no SSN information on CCD records.