National Center for Health Statistics
Data Linkage

The Linkage of the 2016 National Hospital Care Survey to the 2016/2017 National Death Index: Methodology Overview and Analytic Considerations

Data Release Date: January 28, 2019

Document Version Date: January 26, 2022

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 to the 2016/2017 National Death Index: Methodology Overview and Analytic Considerations*, August 2019. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data-linkage/index.htm

Contents

1 Introduction

As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to provide statistical information that can be used to guide actions and policy to improve the health of the American people. 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 August 6, 2019). The 2016 NHCS collected data from a sample of 581 hospitals, of which 158 participated by providing patient-level encounter records. For participating hospitals, these data cover all of their patient ambulatory care and inpatient visits occurring during the course of the year. The NHCS includes detailed information about hospital characteristics, patients' characteristics, and treatment. 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.

Through its data linkage program, NCHS has been able to expand the analytic utility of the data collected from NHCS by augmenting it with mortality data collected from the National Death Index (NDI). This report will describe the linkage of the 2016 NHCS to the 2016/2017 NDI. Although NHCS is not currently nationally representative due to low response rates, 158/581=27%, linking NHCS with the NDI allows for new analyses, such as studying mortality post hospital discharge, along with specific causes of death.

This report includes a brief overview of the data sources, a description of the methods used for linkage, and analytic guidance to assist researchers while using the files. Detailed information on the linkage methodology is provided in the Appendix.

The data linkage work was performed at NCHS under contract #HHSD2002016F92236 by NORC at the University of Chicago with funding from the Department of Health and Human Services' Office of the Secretary Patient Centered Outcomes Research Trust Fund (OS-PCORTF).

2 Background on Linked Files

2.1 National Hospital Care Survey (NHCS)

The NHCS is an establishment survey that collects inpatient (IP), emergency department (ED), and outpatient department (OPD) episode-level data from sampled hospitals. NHCS is one of the National Healthcare Surveys, a family of surveys covering a wide spectrum of healthcare delivery settings from ambulatory and OPD to hospital and long-term care providers. The goal of NHCS, when fully implemented, is to provide reliable and timely healthcare utilization data for hospital-based settings, including prevalence of conditions, health status of patients, health services utilization, and substance-involved ED visits.

From participating hospitals, NHCS collects data on all IP and ambulatory care visits occurring during the calendar year. In previous years of the survey, hospitals were required to provide data from claims records, but to reduce the burden of reporting on participating hospitals, for the 2016 data collection hospitals were given the option of providing their data in the form of electronic health records (EHR) or as claims records. Thus, participating hospitals provided data in the form of Uniform Bill (UB)-04 administrative claim records or EHR data, where the EHR data are an amalgamation of custom extracts and Consolidated Clinical Documents (CCDs). NHCS collects patient PII (e.g., 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 the NDI. The linkage described here includes only IP and ED visits – other, non-ED OPD visits have been excluded.

2.2 National Death Index (NDI)

The NDI is a centralized database of United States death record information on file in state vital statistics offices. Working with these state offices, NCHS established the NDI as a resource to aid epidemiologists and other health and medical investigators with their mortality ascertainment activities.[1] The NDI became operational in 1981 and includes death record information for persons dying in the U.S. or a U.S. territory from 1979 onward.[1] The records, which are compiled annually, include detailed information on the underlying and multiple causes of death.

[1] https://www.cdc.gov/nchs/ndi/index.htm (Accessed August 6, 2019).

3 Linkage Methodology

3.1 Linkage Eligibility Determination

In addition to excluding patients with incomplete PII, the linkage described here only includes patients with at least one IP or ED encounter record reported by NHCS participating hospitals; patients for whom only OPD encounters were reported are excluded from NDI linkage. However, for patients who do have at least one IP or ED encounter in addition to an OPD encounter, the PII from the OPD encounter is utilized, to increase the completeness and accuracy of the hospital record prior to the linkage.

In order for a record to be considered linkage eligible, it must have two of the following: valid date of birth (month, day, and year)[2], name (first, middle, and last)[3], and/or a valid format 9-digit SSN. For example, if the PII on the NHCS record had no SSN, a full name, and only the year of birth, NCHS deems them as being ineligible for linkage, as there would be a considerable likelihood of not being able to find a NDI match even if one exists or a higher likelihood of matching to an incorrect record.

The linkage eligibility status (which indicates whether or not the linkage eligibility criterion is met) for a record is shown by the value of the variable **ELIGSTAT.** The available values include 0 (ineligible) or 1 (eligible). All patients, with at least one IP or ED encounter reported by an NHCS participating hospital, are included on the linked NHCS mortality file.

3.2 Overview of Linkage

The following section outlines steps used to link the 2016 NCHS data to the 2016/2017 NDI. For more details see the Appendix.

The primary identifiers used in the linkage were: SSN, first name, last name, middle initial, month of birth, day of birth, year of birth, death status (according to the discharge status on the hospital record), state of residence, and sex.

The linkage between the 2016 NHCS records and the 2016/2017 NDI records was based on both deterministic and probabilistic approaches. The probabilistic approach performs weighting and link adjudication following the Fellegi-Sunter method.[4] The Fellegi-Sunter paradigm method is the foundational methodology used for record linkage. It estimates the likeliness that each pair is a match before selecting the most probable match between a survey record and NDI record. Following these approaches, a selection process was implemented with the goal of selecting pairs believed to represent the same individual between the data sources. In sum, the steps are the following (to be explained in further detail below):

1. Deterministic linkage, performs joins on exact SSN or the SSN extracted from the Health Insurance Claim Number (HICN)[5] and is validated by comparison of other identifying fields: if

[2] A date of birth is considered to be usable if at least two of the three date parts (year, month, or day) are valid values.

[3] A name is considered to be usable if at least two of these three criteria is met: first name has two or more characters, middle name has one or more characters, and last name has two or more characters.

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

[5] Medicare HICN's contain the beneficiary's SSN. These numbers are actively being phased out by Medicare in favor of Medicare Beneficiary Identifiers (MBIs) https://www.cms.gov/Medicare/New-Medicare-Card/ (Accessed August 6, 2019).

these criteria are met, these records are assigned a probability of being a valid match (match probability) of **1.00**.

2. Probabilistic linkage identifies likely matches, or links, between all records. Records are linked and scored as follows (note: SSN is excluded from the analysis for this step):
    a. Identify possible matched pairs (see Appendix, Section 2.1)
    b. Score potential match pairs—matches are scored based on the concurrence of these variables: First Name, Middle Initial, Last Name (or Father's Surname), State of Residence, Year of Birth, Month of Birth, Day of Birth, Date of Death (if available on hospital record), Sex
    c. Probability modeling – estimate match probability.
3. For each patient record, keep the linked NDI record having the highest estimated match probability as long as it is above the linkage cutoff (see Section 4.7)

The linkage algorithm was developed using SAS 9.4 and was tailored to perform this specific linkage, in order to produce high-quality matches with a low degree of error.

Table 1. 2016 NHCS - Sample Sizes and Unweighted Percentages of Patients Who Were Identified as Deceased in the Interval (2016-2017), by Age

	Sample Size			Percent Deceased	
Age[1]	Total Sample	Eligible for Linkage[2]	Identified as Deceased in 2016 - 2017[3]	Total Sample[4]	Eligible Sample[5]
<18	1,291,571	1,203,905	3,429	0.3%	0.3%
18-44	1,476,724	1,386,108	14,598	1.0%	1.1%
45-64	919,554	864,444	49,531	5.4%	5.7%
>=65	767,846	722,593	144,597	18.8%	20.0%
Total	4,455,695	4,177,050	212,155	4.8%	5.1%

SOURCE: 2016 NHCS linked to 2016/2017 NDI data

NOTES: Data are presented at the patient level. Patients may have multiple encounter records submitted during the survey data collection period but only the last chronological encounter was used in this table.

[1] Age is calculated by subtracting date of birth from the date of discharge from the last chronological encounter record submitted during the survey collection period. Age could not be calculated for 1,367,470 patients due to missing date of birth and are not included in this table. This included 767 patients who matched to the NDI and were identified as deceased.

[2] Eligibility for linkage is based upon having sufficient PII in at least two of three data element groups: SSN, name, and date of birth.

[3] This group includes any patient who was identified as deceased through linkage to the NDI at any time during the linkage interval (2016 - 2017).

[4] This percentage is calculated by dividing the number of deceased patients by the number of patients in the total sample.

[5] This percentage is calculated by dividing the number of deceased patients by the total number of linkage-eligible patients.

4 Analytic Considerations

4.1 NHCS Hospital Eligibility and Sampling

Eligible hospitals for NHCS are non-institutional, non-federal hospitals with six or more staffed IP beds, and there are 6,622 hospitals which met these criteria as of 2013 to form the survey frame. A base sample of 500 hospitals and a reserve sample of 500 additional hospitals was drawn from this frame. Initially, the base sample of 500 hospitals was fielded. In 2013, to provide estimates for substance-involved ED visits, 81 hospitals with 500 staffed IP beds or more were added from the reserve sample. Thus, the hospital sample size for the 2016 NHCS data collection (which re-uses the 2013 sample) was 581 hospitals. In 2016, of the 581 sampled hospitals, 142 hospitals were eligible for linkage (note: this number excludes hospitals that provided records covering less than 6 months of the analysis period). Of those 142 participating hospitals, 140 hospitals sent IP data and 121 hospitals sent ED data.

4.2 Sampling Weights Are Currently Not Available

Currently, there are no sampling weights available for the 2016 NHCS data. This section will be updated if sampling weights are made available in the future. Because the hospital level sampling conducted for the NHCS was not conducted on an equal probability basis, unweighted estimates will be biased to be more similar to those from hospitals selected with higher sampling probability. Similarly, there will be bias towards types of hospitals responding at higher rates. These biases will be more of a concern if estimates vary strongly by factors correlated with sampling and response rates. One way to mitigate these biases in the absence of survey weights is to calculate estimates in the framework of regression modeling that controls for hospital characteristics. This would be done by including hospital characteristics (region, ownership type, and size) as well as patient characteristics (age and sex) among the predictor variables in the model definition. Statistical testing can then be conducted on parameter estimates associated with these characteristics.

4.3 Patient_ID Details

PATIENT_ID is intended to be unique for each individual receiving IP, ED, or OPD services at a participating hospital. However since the de-duplication of patient records required to generate this depends on sometimes incomplete or erroneous data, there may be instances where the same individual is represented by more than one **PATIENT_ID**. This happens infrequently and should not greatly impact analyses. [6]

4.4 Mortality Status

A patient's final determination of vital status can be found using the **MORTSTAT** variable**.** Each patient is assigned a vital status code based on linkage eligibility and mortality status as follows:

- 0 – Eligible for data linkage, assumed alive
- 1 – Eligible for data linkage, assumed deceased based on NDI linkage
- 2 – Eligible for data linkage, assumed deceased from non-NDI source
- 3 – Ineligible for data linkage, assumed deceased from non-NDI source

  . – Ineligible for data linkage, no other source of death available

[6] For more information of Patient_ID generation, see Technical Notes on page 14: https://www.cdc.gov/nchs/data/nhsr/nhsr097.pdf (Accessed August 6, 2019).

4.5 Mortality Source Information

The source of the death information is indicated by two variables:

- National Death Index (**MRT_SOURCE_NDI**) – Result of deterministic/probabilistic linkage to NDI.
- Data Collection/Hospital Record (**MRT_SOURCE_DCL**) – Result of survey data collection, patient was discharged dead from the hospital.

For each mortality source variable, a value of 1 indicates the patient is deceased, while a numeric missing value (.) indicates the patient was either not deceased, the source date-of-death did not match the NDI date-of-death (when linked to the NDI), or was not eligible for linkage. When more than one mortality source variable indicates the patient is deceased (each take a value of 1) then the date-of-death is matching between the sources. For example, if a patient was identified as deceased by means of the linkage to the NDI and by a hospital discharge code and both date-of-death were the same, then both **MORTSRCE** variables (NDI and DCL) will have a value of one. If the same scenario occurs but the date-of-death are different, **MORTSRCE_NDI** will receive a value of one and **MORTSRCE_DCL** will be set to numeric null. These variables should be used only for informational purposes. Please see Section 4.4 Mortality Status for more information on using **MORTSTAT** to restrict the linked data by vital status.

4.6 Linkage of Patient Records with Improbable Ages & Multiple Dates of Birth

The 2016 NHCS-NDI linked mortality file includes records where the calculated age presumed alive at the end of mortality follow-up is either negative or 110 years or more. Given the probabilistic nature of the mortality ascertainment and the lower likelihood of being alive at 110 years or older, analysts may wish to consider these cases as lost to follow-up and make them ineligible for mortality analyses. (Note: NDI only includes deaths that occurred in the United States or a U.S. territory and therefore may not include deaths of all patients). For individuals with negative ages, the survey reported date of birth occurs after the end of the 2016 year of data collection. Analysts may also wish to consider these post survey collection dates of birth as reporting errors and exclude these individuals from mortality analysis.

A practical method for determining an age cutoff at which patients should be considered lost to follow-up is to use the probability of a member in a particular population dying at, or living to, a particular age. The Social Security Administration (SSA) published a report in 2005 (Life Tables for the United States Social Security Area 1900-2100. SSA Pub No. 11-11536) containing projections of mortality for cohorts of births in decennial years 1900 through 2100. Based on these cohort life tables the NCHS Data Linkage Team calculated probabilities of death, conditional on year of birth and sex, but not adjusted for last known alive year (typically the year of data collection). These probabilities are available for researchers upon request (datalinkage@cdc.gov). Please refer to the SSA report (https://www.ssa.gov/OACT/NOTES/pdf_studies/study120.pdf (Accessed August 6, 2019) for more information.

The 2016 NHCS analytic file can potentially contain invalid or multiple dates of birth. An invalid date of birth is defined as a date of birth that does not follow conventional date structure (e.g., having a day outside of 1 – 31) or a date of birth that occurs after the end of the 2016 year of data collection period. As date of birth may be reported on each hospital encounter record, reporting errors may occur resulting in multiple dates of birth collected for the same patient. Researchers using the 2016 NHCS analytic file linked to the 2016/2017 NDI file, should consider using the adjudicated date of birth fields (**DOBDAY**, **DOBMONTH**, and **DOBYEAR**) for analyses utilizing date of birth.

4.7 Restricted-Use Linked Mortality File Linkage Results Variables

Data linkages include some uncertainty over which pairs represent true matches. For the 2016 NHCS data linked to the NDI, we set the probabilistic cut-offs for determining which pairs were considered a link (an inferred match) at those values that minimized the sum of our estimated counts of Type I error (false positive links— identified as deceased but actually alive) and Type II error (false negative non-links—identified as alive but actually deceased). However, for each candidate pair, we computed a probability of match validity **(PROBVALID)** based on the total pair weight **(PAIRWGT)** (see Appendix Sections 2.2 and 2.3 for a discussion of how these values were computed). Researchers can access and use these variables, upon approved requests, to adjust linkage certainty or for sensitivity analyses of vital status.

In the file, we used a **PROBVALID** value cutoff of > **0.9525** for the 2016 NHCS data, which is the threshold that produces the lowest total error – both Type I and Type II. In their RDC proposal, researchers may request **PROBVALID** and change the link acceptance cut-off. For some analyses, it may be useful to estimate the sensitivity of derived estimates to pairs with lower probability of being valid matches by lowering the link acceptance criteria **PROBVALID** values to below the cutoffs. For other analyses, it may be desirable to minimize Type I error, which would be the result of a value of **PROBVALID** closer to **1.0000**.

In addition, individual agreement weight (pair weights components) variables are available to researchers that indicate the components of the matching variables. The total pair weight, **PAIRWGT**, is the sum of nine pair weight components:

- First Name or First Initial
- Middle Initial
- Last Name/Father's Surname *(Note: These are two separately computed pair weights. However, only highest one of the two is used in the tabulation of the final pair weight.)*
- Year of Birth
- Month of Birth
- Day of Birth
- Sex
- State of Residence
- Date of Death/Discharge Date

Each PAIRWGT represents a specific identifier comparison.[7] These component values are also available to researchers upon request in their RDC proposal. For more information on how the total pair weights are calculated, refer to the methodology Section 2.2 Score Possible Matched Pairs. When looking at the 9 component pair weights simultaneously, the one having the highest positive value shows the identifier agreement most indicative of being a match and the one having the lowest negative value shows the identifier non-agreement most indicative of not being a match. Within the linked files, variables are provided for all patients who returned a potential NDI match, independent of whether the patient's status was assumed deceased by the algorithm. These variables allow researchers to assess the linkage results for patients whose final status was assumed to be alive and to conduct sensitivity analyses among decedents. The complete list of variables is contained in the Death Certificate and NDI Match

[7] Find details for each **PAIRWGT** variable in data dictionary posted on this linkage's website at https://www.cdc.gov/nchs/data-linkage/nhcs-ndi.htm (Accessed August 6, 2019).

Variables Data dictionary: https://www.cdc.gov/nchs/data-linkage/nhcs-ndi.htm (Accessed August 6, 2019).

5 Access to Data Files

5.1 Access to the Restricted-Use NHCS-NDI Linked Mortality File

To ensure confidentiality of data, NCHS provides safeguards including the removal of all personal identifiers from analytic files. Additionally, the files containing these linked data are only made available in secure facilities for approved research projects. Researchers who wish to obtain access to the linked 2016 NHCS to 2016/2017 NDI file must submit and have an approved research proposal to the NCHS Research Data Center (RDC): https://www.cdc.gov/rdc/index.htm (Accessed August 6, 2019). The proposal will be evaluated for feasibility and disclosure risk. NCHS RDCs are housed on-site at Centers for Disease Control and Prevention (CDC) facilities in Hyattsville, MD, Washington, D.C., and Atlanta, GA. In addition, NCHS restricted data can be accessed from RDCs housed in U.S. Census Bureau offices in several locations across the country. Researchers generally will need to be on-site at one of the RDCs to access restricted-use linked data, including the restricted-use NHCS-NDI-linked files. Within the RDC, the NHCS-NDI linked files can be merged with NHCS restricted use survey data files using a unique de-identified patient identification numbers.

5.2 Combining the Linked NHCS-NDI File to NHCS Analytic Files

NHCS is an establishment survey where the respondents are individual hospitals rather than their patients. Typically this type of survey restricts analyses to the sample unit-level, but because NHCS collects hospital encounter-level records, encounter-level analysis is also possible. For each patient with either an IP discharge or ED visit, results of the person-level linkage to the NDI are available in the NHCS-NDI linked mortality files.

To perform encounter-level analysis, the NHCS-NDI linked mortality file can be used in conjunction with NHCS analytic files[8], which are also available within the NCHS RDC. The linked mortality file includes variables such as Patient ID, date of birth, date of death, and cause of death information, while the analytic files include analytically-pertinent hospital-level details (such as bed size and geographic region) and episode-level details (patient demographics, diagnoses, procedures, admission and discharge dates). To integrate the analytic file details into the NHCS-NDI linked mortality file, joins should be made on the common field, **PATIENT_ID**.

5.3 Death Certificate Information

Additional data, obtained from the death certificate, are available to researchers using the restricted-use 2016 NHCS-2016/2017 NDI linked mortality file. These variables are prefixed with DVS * and are populated for different years of death year. The data dictionary, Death Certificate and NDI Match Variables, on the Restricted Use Linked 2016 NHCS- 2016/2017 NDI Data webpage contains the complete list of variable names, labels, and other meta-data as described in Section 4.7: https://www.cdc.gov/nchs/data-linkage/nhcs-ndi.htm (Accessed August 6, 2019). If more information (e.g., definition of values) is sought about these variables, please refer to the NVSS Public Use Data File

[8] Find more information about the NHCS analytic file: https://www.cdc.gov/rdc/b1datatype/dt1224h.htm (Accessed August 6, 2019).

Documentation webpage: https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm (Accessed August 6, 2019).

Appendix: Detailed Description of the Linkage Methodology

1 Deterministic Linkage Using SSN

The first step in the linkage process is to attempt a deterministic linkage for all eligible NHCS records that were submitted with a valid format SSN or an SSN extracted from a HICN. The deterministic linkages were validated by comparing first name, middle initial, last name, month of birth, day of birth, year of birth, and state of residence identifiers in order to ensure that the records were a valid match. If the ratio of matching identifiers to non-missing identifiers is greater than 50%, the linked pair is retained as a deterministic match. The collection of records resulting from the deterministic match is referred to as the 'truth deck.'

2 Probabilistic Linkage

In order to infer that a pair is a match, the linkage algorithm first identifies potential match pairs (links) and then evaluates their probable validity (i.e., that they do represent the same individual). The following sections describe these steps in detail. This linkage methodology closely follows the Fellegi-Sunter paradigm method, the foundational methodology used for record linkage. The method estimates the likelihood that each pair is a match – using formulaic pair weights computed for each identifier in the pair – before selecting the most probable match between two records.

2.1 Identify Possible Matched Pairs

The first step in the probabilistic matching process is to identify possible matched pairs between the records representing individual persons from both files. This method defines a **possible matched pair** as: 1) A set of two records with at least three of four matching identifiers (listed below) or 2) Having chance agreement below a minimal probabilistic threshold (described below). The first method identifies the possible matched pairs, by creating a Cartesian product between eligible NHCS records and NDI records. The Cartesian product includes every possible combination of records (i.e., each NCHS record entering this analysis is virtually compared to all NDI records). The sum of the following four identifiers (this value is stored in the variable **CC** (for Comparison Count), which can be as high as 4, is used to identify possible matches:

- First Name – NYSIIS Coding[9]
- Last Name or Father's Surname – NYSIIS Coding[9]
- Month and Date of Birth (Combined field)[10]
- Year of Birth and State of Residence (Combined field)

If CC ≥ 3 the records are classified as a possible match.

The second method to identify possible matched pairs, uses the Cartesian product to compute the expected count of *spurious* NHCS records paired to an NDI record:

1. For each identifier analyzed on each NDI record, the proportion of NHCS records that share that identifier field is computed. For example, for first name Mary, 0.864% of NHCS records are determined to share the first name Mary, and so the probability of a spurious match to an NDI

[9] NYSIIS coding is a software used to identify possible alternative spellings of first and last names provided in data files. Using this software increases the likelihood of matching true pairs that might have been missed from either spelling errors or nicknames used in the original fields provided.

[10] The linkage uses a combined field to speed processing.

record with first name Mary is assigned that value. That same proportion is computed for all values for all four identifiers.

2. The proportions for each agreeing identifier are multiplied together to estimate the probability of spurious agreement.
3. The record-level proportion is multiplied by 350 million[11] to approximate the total number of people at risk of hospitalization in the United States in 2016. The result is a naïve Bayes estimate.[12] In mathematical terms, the naïve Bayes estimate is $\prod_i P_i$ where the P_i are the marginal probabilities of each identifier considered. Here, it estimates the expected number of spurious NHCS records paired to an NDI record in the Cartesian product. In sum:

   Expected Spurious Matched Records=
   350,000,000 × (First Name proportion × Last Name/Surname proportion ×
   Month and DOB proportion × Year of birth and state of residence proportion)

If the *expected spurious matched records* estimate is less than 1 then this is probably not a spurious agreement. To broaden the pairs under consideration, the linkage accepts patient records with an expected spurious match score 2 or less. Therefore, the threshold is defined as a score of 2. If this level is not exceeded, the pair continues for further evaluation (described starting in Section 2.2). Pairs that exceed this threshold are excluded from further evaluation because they are not considered possible matches.

Table 1 explains the theoretical proportions (estimates) and calculations for a pair consisting of an NDI record and NHCS record agreeing on two identifiers (First Name, Last Name), but not on Month/Date of Birth or Year of Birth/State of Residence. In this example the first name and last name are rare and have a low probability in the population of being a match (e.g., recall that Mary had a high value for first name 0.864%).

Table 1: Expected Spurious Matched Records Calculation Example

Identifier	Value
First Name	0.00015
Last Name	0.000035
Month/Date of Birth	N/A[13]
Year of Birth/State of Residence	N/A[13]
Record-Level Result	0.00015 x 0.000035 = 5.25E-9
Exp. # Spurious Records	5.25E-9 x 350,000,000= 1.8375

For this example, the expected spurious NHCS records estimate is 1.8375, which qualifies the pair as a possible matched pair since the estimate is less than two. Note that it is only because there was

[11] Annual Estimates of the Resident Population: April 1, 2010 to July 1, 2017 Source: U.S. Census Bureau, Population Division.

[12] By naïve-Bayes, the linkage assumes that the conditional probabilities are equal to the marginal probabilities.

[13] Since there is no agreement on these fields do not enter the computation.

agreement on rare first and last names (as corresponds to the low agreement probabilities) that this pair met the qualifying threshold.

2.2 Score Possible Matched Pairs

According to Christensen, "The more similar values two records have in common across these attributes, the more likely it will be that they correspond to the same individual."[14] After identifying possible candidate pairs, each possible matched pair (those records with at least three linking identifiers, **CC** ≥ 3 in agreement or with an expected spurious match count less than two) is 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 are used in a probability model (explained in Section 2.2.4), which allows the linkage to select final pairs to include in the linked file. The scoring process follows the following order:

2.2.1 Calculate M- and U- probabilities

2.2.2 M and U probabilities for First and Last names

2.2.3 Calculate agreement and non-agreement weights

2.2.4 Calculate pair weight scores

The scoring of possible matched pairs is calculated on the following identifiers:

- First Name or First Initial
- Middle Initial
- Last Name/Father's Surname *(Note: These are two separately computed pair weights. However, only highest one of the two is used in the tabulation of the final pair weight.)*
- Year-of-Birth
- Month-of-Birth
- Day-of-Birth
- Sex
- State-of-Residence
- Date-of-Death/Discharge-Date

2.2.1 Calculate M- and U- Probabilities

For each of the nine identifiers listed above, the linkage computes the **M-probability** – the probability that the identifiers from the records in question agree, given that the two records are a match. It computes these M-probabilities from the validated pairs in the truth deck. For example, among the validated pairs, 99.4% agree on year of birth and 99.7% agree on state of residence. Therefore, the M probability of year of birth is 0.994 and state of residence is 0.997. The same process is computed and used for each of the 9 identifiers. These percentages are then used as the M-probabilities.

The **U-probability** – the probability that the two values for an identifier from paired records agree, given that they are NOT a match – is computed specific to each individual value for a given linkage variable.

[14] Christen, Peter. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Berlin Heidelberg: Springer-Verlag, 2012. http://www.springer.com/us/book/9783642311635 (Accessed August 6, 2019).

For example, the U-probability of state of residence for a state with many patients would be around is 0.06 (6.0%) but for a state with less patients the U-probability would only be about 0.0003 (0.03%) because records from individuals residing in that state are less common in the data file. The U-probabilities are based on the frequency of the data element in the NHCS data file and are computed from a frequency tabulation on the full set of NHCS records in the submission file.

2.2.2 M and U Probabilities for First and Last Names

Similar to the M-probabilities, Jaro-Winkler levels (85, 90, 95, and 100) are also calculated for use in the U-probability section. The manner of their creation is identical to the process described above. For several reasons, the first and last name U probabilities were computed differently than for the remaining comparison variables. Due to the many possible values of first and last names, it would be impractical to compute a U-probability for each individual name value. Only the names that appear in the submission file will be used in the scoring process and will have a U-probability calculated, all other names are excluded during this process. Of note, names of one character or fewer (missing) will not be included in this calculation. Names with length of one will be used to compare name initials, when present.

Complete name tallies (separately, for first and last names) were produced for each NHCS name present in the submission file. For each level of name on the file, we compare all other names present in the submission file. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. For each of the NHCS names present in the submission file, the number of submission entries that have a name agreeing at each of the Jaro-Winkler levels were tallied.[15,16,17] The U-probability is then calculated by dividing the tally count by the total number of records (where a name has more than one character).

2.2.3 Calculate Agreement and Non-Agreement Weights

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

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

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

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

2.2.4 Calculate Pair Weight Scores

The next step is to calculate pair weights, which are used in the probability model. The pair weights are the summation of the identifier-specific agreement and non-agreement weights. Therefore, if the

[15] 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.

[16] 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.

[17] Winkler, William E. Frequency-based matching in Fellegi-Sunter model of record linkage. Bureau of the Census Statistical Research Division 14. 2000.

- Identifier agrees: Add identifier-specific agreement weight into pair weight
- Identifier disagrees: Add identifier-specific non-agreement weight (negative value) into pair weight
- Identifiers cannot be compared because one or both identifiers from the respective records compared are missing: no adjustment made to the pair weight

Once each individual pair weight has been calculated, a full pair weight is created from the summation of the individual weights. The full pair weights are then used to develop a probability model by the linkage algorithm.

2.3 Probability Modeling

A probability model, developed from a logistic regression analysis, estimates the likely validity that each pair is a match. Those pair weight scores allows the linkage algorithm (as described in Section 2.2.4) to first determine which pair to select when multiple pairs are available for a given patient ID, and then determine whether the pair's likely validity is great enough to keep in the final set of accepted matches. Since multiple pairs may exist for the same patient, the probability calculated with the model also allows the linkage to identify the pair with the highest probability of being a match.

The logistic regression model was estimated on possible matched pairs that had a valid and formatted SSN on both the NHCS claim record and the NDI record. Recall, SSN was not used as a match identifier. The response variable used in the regression analysis is the agreement on SSN, created as a categorical variable (1 – SSN values agree, 0 – SSN values do not agree).[18] The regression model estimates the likely validity that each pair is a match given its pair weight. Since multiple pairs may exist for the same patient, the model allows the linkage algorithm to calculate the probability of validity for each possible pair, which can then be used to identify the pair with the highest probability of being a valid match.

There are two variables that enter into a logistic regression model:

- The number of identifiers in agreement from among those in the Select Possible Matched Pairs section: <**CC>** -- (see discussion in Section 2.1)
- Pair weight

The resulting model estimates the probability of match validity.

3 Select Matches for Final File

Up to this point, the linkage has identified possible matches through both the deterministic linkage and the probabilistic linkage. These identified matches all have a probability value assigned that measures their probability of being a valid match. The deterministic matches were automatically assigned a probability value of (1), while the probabilistic links were assigned a probability of validity[19] using the logistic model.

The penultimate step is to assign a probability threshold that a pair is a valid match. This probability threshold has been set to 0.9525 for all records. The threshold was set at the level that produced the lowest estimated total error (Type I and Type II). If the best possible match has an estimated validity less than the threshold, then the linkage algorithm will not accept it into the final matched file.

[18] The linkage classifies it as an agreement if five or more of the nine SSN positions have the same digit on the SSN values being compared.

[19] The probabilistic linkage match validity is estimated by logistic regression, value between zero and 1 (non-inclusive).

Last, the linkage algorithm selects only one pair per patient on the NHCS file – of those pairs that met the probability thresholds just discussed – to include in the final matched file. If there is only one possible pair for a given patient above the relevant threshold, then that pair is included in the final file. If there is more than one possible matched pair for a given patient above the relevant threshold, then the possible matched pair with the highest probability of being a valid match is selected. If a tie remains at this point then the record with the better matching information is selected. If all information is matching the same, one record is selected at random. Note – if one of these possible pairs was created during the deterministic linkage, then this pair will always be selected because it is assigned a probability of one. For this linkage the estimated Type I error was 0.24% and the estimated Type II error was 1.14%.