National Center for Health Statistics
Data Linkage

# The Linkage of the 2014 National Hospital Care Survey to the 2014/2015 Centers for Medicare & Medicaid Services Master Beneficiary Summary File: Methodology Overview and Analytic Considerations

Data Release Date: January 28, 2019

Document Version Date: August 6, 2019

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 2014 National Hospital Care Survey to the 2014/2015 Centers for Medicare & Medicaid Services Master Beneficiary Summary File: 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 2014 NHCS collected data from a sample of 581 hospitals, of which 95 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 healthcare services data collected from the Centers for Medicare & Medicaid Services' (CMS) Medicare Master Beneficiary Summary File (MBSF). This report will describe the linkage of the 2014 NHCS to the 2014/2015 CMS MBSF. Although NHCS is not currently nationally representative due to low response rates, 95/581=16%, linking NHCS with the CMS data allows for new analyses, such as examining comorbidities and utilization of non-inpatient related health care services.

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

The National Hospital Care Survey 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 Health Care Surveys, a family of surveys covering a wide spectrum of healthcare delivery settings from ambulatory and outpatient 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. Participating hospitals provide these data in the form of Uniform Bill (UB)-04 administrative claim records. Unlike its predecessor surveys, NHCS also collects PII (e.g., name, date of birth, and Social Security Number (SSN)), which allows for the linkage of episodes of care across hospital units and to other data sources, such as the CMS Medicare MBSF. The linkage described in this document includes only IP and ED visits – other, non-ED OPD visits have been excluded.

## 2.2 Centers for Medicare & Medicaid Services, Master Beneficiary Summary File

The CMS Medicare MBSF is an annual file containing demographic and enrollment information about beneficiaries enrolled in Medicare during each calendar year. The CMS Medicare MBSF includes information on beneficiary demographic characteristics, reason for Medicare entitlement, and program enrollment type (Original Medicare vs. Medicare Advantage (MA)).

The **Base (A/B) segment** includes beneficiary characteristics, monthly entitlement indicators, reasons for entitlement (initial and current), and monthly Medicare Advantage indicators. The **Part D segment** includes variables specific to Medicare Part D Prescription Drug plan enrollment. The **Cost & Utilization segment** includes summarized information about the service utilization and Medicare payment amounts by type of claim, including prescription drugs. The **Chronic Conditions segment** includes variables that indicate a Medicare beneficiary has received a service or treatment for selected chronic health conditions.[1]

[1] Conditions Included in CCW: acquired hypothyroidism, acute myocardial infarction, Alzheimer's Disease, Alzheimer's Disease & related disorders or senile dementia, anemia, asthma, atrial fibrillation, benign prostatic hyperplasia, cancer (colorectal), cancer (endometrial), cancer (female/male breast), cancer (lung), cancer (prostate), cataract, chronic kidney disease, chronic obstructive pulmonary disease (COPD) and bronchiectasis, depression, diabetes, glaucoma, heart failure, hip / pelvic fracture, hyperlipidemia, hypertension, ischemic heart disease, osteoporosis, rheumatoid arthritis / osteoarthritis, stroke / transient ischemic attack

# 3 Linkage Methodology

## 3.1 Linkage Eligibility Determination

Linkages were only possible for NHCS patient records that had certain minimum levels of PII available. In order for a record to be considered linkage eligible, the record must have valid date of birth (month, day, and year)[2] and name (first, middle, and last)[3] information present. For example, if the PII on the NHCS claims records had a full name and only the year of birth, they were deemed as being ineligible for linkage, as there would be a considerable likelihood of not being able to find a CMS match even if one exists.

The linkage eligibility status (which indicates whether or not the linkage eligibility criterion has been met) for a record can be ascertained using the variable **ELIGSTAT.** The available values include 0 (ineligible) or 1 (eligible). Of note, only eligible patients that matched to a CMS enrollment record are included on the linked NHCS – CMS Medicare MBSF file.

## 3.2 Overview of Linkage

The following section outlines steps used to link the 2014 NHCS data with the 2014/2015 CMS MBSF. For more details see the Appendix.

The primary identifiers used in the linkage were: SSN, Medicare Health Insurance Claim Number (HICN), first name, last name, middle initial, month of birth, day of birth, year of birth, zip code of residence, state of residence, and sex. Corresponding CMS identification data is stored in the CMS Enrollment Database (EDB). NHCS patient records are linked using the CMS EDB.

The linkage between the 2014 NHCS records and the CMS EDB was based on both deterministic and probabilistic approaches. The probabilistic approach performs weighting and link adjudication as described in the Fellegi-Sunter paradigm method.[4] Following these methods, a selection process was implemented with the goal of selecting pairs believed to represent the same individual between the data sources. Table 1 highlights the linkage results. In sum, the linkage steps are the following (to be explained in further detail in the appendix):

1. Deterministic linkage, performs joins on exact SSN or HICN and is validated by comparison of other identifying fields
2. Probabilistic linkage identifies likely matches, or links, between all records. If a deterministic match exists it is assigned a probability of 1, other records are linked and scored as follows:
    a. Identify possible matched pairs via blocking
    b. Score potential match pairs
    c. Probability modeling – assign probability that pairs are matches
3. Select pairs believed to represent the same individual between data sources

[2] A date of birth is considered to be usable if at least two of the three date parts are valid date 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.

All 2104 NHCS records successfully matched to the CMS EDB are then extracted from the 2014-2015 CMS Medicare MBSF.

**Table 1. Linked 2014 NHCS – 2014/2015 CMS Medicare MBSF - Sample Sizes and Percent Linked, by Age**

| | Sample Size | | | Percent Linked | |
|---|---|---|---|---|---|
| | Total Sample | Eligible for Linkage[2] | Linked to 2014-2015 Medicare Administrative Data[3] | Total Sample[4] | Eligible Sample[5] |
| **Age[1]** | | | | | |
| <65 | 2,946,281 | 2,685,538 | 234,527 | 8.0% | 8.7% |
| >=65 | 610,784 | 550,221 | 538,451 | 88.2% | 97.9% |
| Total | 3,557,065 | 3,235,759 | 772,978 | 21.7% | 23.9% |

NOTES: Data are presented at patient level. Patients were chosen by selecting the last chronological record within the survey timeframe. Age could not be determined for 1,221 patients based on available data and they are not included in this table.

[1] Age is based on the survey participant's assumed age at final encounter (date of last known contact).

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

[3] This group includes linkage-eligible patients who linked to Medicare MBSF administrative records at any time during the linkage interval (2014 - 2015).

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

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

# 4 Analytic Considerations

This section summarizes some key analytic issues for users of the linked NHCS data and CMS administrative records. It is not an exhaustive list of the analytic issues that researchers may encounter while using the linked 2014 NHCS-2014/2015 CMS Medicare MBSF. 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). Users of the linked Medicare data are encouraged to visit the ResDAC website http://www.resdac.org (accessed August 6, 2019) for more information on Medicare data.

## 4.1 Sampling Weights Are Currently Not Available

Currently, there are no sampling weights available for the 2014 NHCS data. This section will be updates if sampling weights are made available in the future. Because the hospital level sampling conducted for 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 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.

## 4.2 Hospital Linkage Eligibility

While most participating hospitals provided a substantial majority of patient records with sufficient information to be linkage eligible, for some hospitals, most or all records omitted the required data fields for linkage eligibility. Analysts may wish to exclude all patient records from these hospitals when analyzing linked CMS data. The linkage eligibility distribution for each hospital can be reviewed by cross tabulating hospital ID, **HOSPID** (from the NHCS analytic files), with linkage eligibility status, **ELIGSTAT**. Ninety one percent of the patients from participating hospitals were considered to be eligible for the linkage to CMS.

## 4.3 Patient_ID Details

**PATIENT_ID** is a de-identified ID that 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 ID 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.[5]

## 4.4 Medicare Advantage

CMS generally does not receive fee-for-service claims for Medicare beneficiaries who are enrolled in Medicare Advantage (including private fee-for-service plans paid on a capitation

[5] 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)

basis). Medicare Advantage plans are also referred to as Medicare Part C and include Health Maintenance Organizations (HMOs), Preferred Provider Organizations (PPOs), Private Fee-for-Service (PFFS) Plans, Special Needs Plans, and Medicare Medical Savings Account Plans. During the time covered by the linked data files, Medicare Advantage enrollment reached 31% of total Medicare beneficiaries.

Researchers should consider the percent of participants enrolled in a Medicare Advantage program when determining the feasibility and sample sizes of their proposed research projects. Medicare Advantage enrollment can be identified using the HMO indicators from CMS Medicare MBSF – Part A/B Segment. The file includes 12 HMO indicator variables (HMO_IND_01-HMO_IND_12), one for each month. During periods of Medicare Advantage enrollment, beneficiaries do not generate claims when using Medicare-covered services, except for selected services. Enrollees in cost-based plans may also generate some claims for IP hospital services. Utilization of most Medicare-covered services is unobservable from Medicare claims data during periods of Medicare Advantage enrollment. Therefore, in general, studies based on analysis of claims data should exclude Medicare Advantage enrollees from their beneficiary samples.

For more information on how to create an analytic sample that excludes Medicare beneficiaries enrolled in a Medicare Advantage plan, refer to a document written by ResDAC https://www.resdac.org/articles/identifying-medicare-managed-care-beneficiaries-master-beneficiary-summary-or-denominator (accessed August 6, 2019) or contact ResDAC, which provides free consultation for researchers using Medicare files, http://www.resdac.org (accessed August 6, 2019).

## 4.5 Cost Sharing

Medicare beneficiaries often have a number of cost sharing requirements (i.e. deductibles and coinsurance). Although claims are generated for services where beneficiary cost sharing is involved, the Medicare payment amount does not necessarily represent the full cost to the beneficiary for the service. It is not possible to determine whether the beneficiary paid the cost-sharing amount “out-of-pocket” or whether the cost-sharing was paid by a third party, such as Medi-gap. Therefore, the total amount spent for a given healthcare service may not be captured by relying on the claims data alone.

## 4.6 Medicare Payment and Conditions Data

The CMS Medicare MBSF Cost and Utilization segment includes one record for each beneficiary enrolled in Medicare in the calendar year of the file. This record includes summary utilization and total annual payment for Medicare covered services including hospitalizations and physician visits. The CMS Medicare MBSF variables associated with costs and payments may contain extreme outliers. Users may wish to consider applying top or bottom coding limits for these variables as these extreme values may adversely affect statistical calculations. Additional information about the variables included in the CMS Medicare MBSF Cost and Utilization segment is available at https://www.resdac.org/cms-data/files/mbsf-cost-and-utilization (accessed August 6, 2019).

The CMS Medicare MBSF Chronic Conditions segment flags each Medicare beneficiary for the presence of one of 27 specific chronic conditions. Additional information about the methodology used to assign chronic condition flags to Medicare beneficiaries is available at

https://www.ccwdata.org/web/guest/condition-categories (accessed August 6, 2019). CMS cautions users that it is not possible to attribute summary utilization or payment data to a given specific chronic condition as beneficiaries may have other health conditions that contribute to their annual Medicare utilization and payment amounts. (https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/Downloads/Methods_Overview.pdf, accessed August 6, 2019)

## 4.7 Utilizing Administrative Race Data

While a race variable field is available from hospitals that submitted UB-04 claims, the percent of patients with a valid race code is low (only about 10% have valid codes). Researchers may wish to consider utilizing the race and ethnicity data present in the linked CMS administrative records. The CMS Medicare MBSF provides two race and ethnicity variables **BENE_RACE_CD** (accessed August 6, 2019) and **RTI_RACE_CD** (accessed August 6, 2019) located in the A/B Segment. BENE_RACE_CD is the variable reported in the CMS administrative claims data system. The variable RTI_RACE_CD contains race and ethnicity codes imputed through the use of an algorithm developed by the Research Triangle Institute (RTI) and used by CMS to improve the accuracy of race and ethnicity data reported in the administrative claims data system. More detailed information regarding the RTI algorithm can be found at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4195038/ (accessed August 6, 2019).

## 4.8 On CMS Medicare MBSF Records with No Claims Data

There may be instances where a patient in the NHCS is linked to a CMS Medicare MBSF record, but no claims data are available. It is possible to be enrolled in Medicare but not utilize Medicare services during the coverage period. In addition, there may be some record keeping inconsistencies because CMS data are collected for administrative, not research purposes.

## 4.9 Medicare Entitlement Variables

The CMS Medicare MBSF includes three variables indicating Medicare entitlement: original reason for entitlement, current reason for entitlement, and Medicare status code.

These variables can be located in the Base A/B segment table. A beneficiary's *original reason* for Medicare entitlement is found in the variable ENTLMT_RSN_ORIG. This variable is coded by CMS using information provided by the Social Security Administration and/or Railroad Retirement Board. Knowing a beneficiary's original reason for entitlement can be useful for identifying which aged beneficiaries were formerly Medicare disabled, since their cost and utilization profiles tend to differ from other aged beneficiaries, especially at ages 65-74. ENTLMT_RSN_ORIG values include: Old Age and Survivors Insurance (OASI), Disability Insurance Benefits (DIB) and End Stage Renal Disease (ESRD).

A beneficiary's *current reason* for Medicare entitlement is found in the variable ENTLMT_RSN_CURR. Possible values include: OASI, DIB and ESRD.

The variables MDCR_STATUS_CODE_01 - MDCR_STATUS_CODE_12 specify the monthly status of the beneficiary's entitlement to Medicare benefits. Possible values include: Aged without ESRD, Aged with ESRD, Disabled without ESRD, Disabled with ESRD, and ESRD only.

### 4.10 File Year Indicator

The reference year can be found in the variable BENE_ENROLLMT_REF_YR. Please note that both 2014 and 2015 records are stacked in this variable. It is possible that a single beneficiary can have records for both 2014 and 2015. If this is the case, the beneficiary will appear twice in the file.

# 5 Access to Data Files

## 5.0 Access to the Restricted-Use Linked NHCS – CMS Medicare MBSF

To ensure confidentiality of health data, NCHS provides safeguards including the removal of all personal identifiers from analytic files. Additionally, the files containing these data are only made available in secure facilities for approved research projects. Researchers who want to obtain the linked 2014 NCHS- 2014/2015 CMS Medicare MBSF files must submit a research proposal to the NCHS Research Data Center (RDC) to obtain permission to access the restricted use files. All researchers must submit a research proposal to determine if their project is feasible and to gain access to these restricted data files. The proposal provides a framework which allows RDC staff to identify potential disclosure risks. More information regarding RDC and instructions for submitting an RDC proposal are available from: https://www.cdc.gov/rdc/ (Accessed August 6, 2019).

## 5.1 Combining the Linked NHCS-CMS Medicare MBSF to NHCS Analytic Files and Linked NDI data

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 CMS Medicare MBSF are available in the linked 2014 NHCS-2014/2015 CMS Medicare MBSF.

To perform encounter-level analysis, the linked 2014 NHCS-2014/2015 CMS Medicare MBSF files can be used in conjunction with 2014 NHCS analytic files[6] and the linked 2014 NHCS-2014/2015 NDI mortality file, which are also available through the NCHS RDC. The linked mortality file includes 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 NHCS analytic and the linked 2014 NHCS-2014/2015 mortality file into the linked 2014 NHCS- 2014/2015 CMS Medicare MBSF, joins should be made on the common field, **PATIENT_ID**. Additionally, **PATIENT_ID** allows linkage of multiple visits for the same patients within or across hospital settings (IP or ED).

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

# Appendix I: Detailed Description of Linkage Methodology

## 1 Deterministic Linkage Using Unique Identifiers

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 HICN. In some cases, the SSN field does not hold a validly formatted value and a possible replacement value using the HICN is instead used.[7] Using HICN to derive a SSN when no valid SSN was available resulted in a 13.35% increase IP ID's with a valid SSN.

The deterministic linkages were validated 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 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, and that it estimates the likeliness 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 Blocking

Blocking is a key step in record linkage. It identifies potential candidate pairs without comparing every single pair in the Cartesian product. According to Christen, blocking or indexing, "splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key)."[8] Rules can be used to define the blocking criteria however, for this linkage, instead of rules, we used the data to help create a set of blocks that would efficiently join the datasets together. By using the data to create the efficient block set, we ideally reduce the number of false positive links while retaining a high percentage of true positive links. For the purpose of this linkage, the 'truth deck' was used as the training dataset. When the data are used in this manner, it is commonly referred to as a machine learning algorithm. For more detailed information on the method that was used please refer to "Learning Blocking Schemes for Record Linkage."[9]

---

[7] Only the HICN's where the individual was listed as the primary beneficiary were used as replacements for SSNs

[8] 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).

[9] Michelson, Matthew, and Craig A. Knoblock. "Learning Blocking Schemes for Record Linkage." In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, 440–445. AAAI'06. Boston, Massachusetts: AAAI Press, 2006. https://pdfs.semanticscholar.org/18ee/d721845dd876c769c1fd2d967c04f3a6eeaa.pdf (Accessed August 6, 2019).

Prior to the probabilistic linkage, a machine learning algorithm was applied to the data with the goal of creating a blocking scheme. Utilizing the 'truth deck' and subsets of the survey and CMS records, the algorithm generated 7 blocks to be used in the blocking scheme. The blocking scheme is designed to reduce the number of results generated in the Cartesian product and decrease overall workload. Additionally, the algorithm is designed to minimize the number of missed true match pairs during this process. Table 1 provides a specific breakdown of each block by variable.

**Table 1. Breakdown of block variables and scoring variables used to identify and score linked records**

| Block 1 | Block 2 | Block 3 | Block 4 | Block 5 | Block 6 | Block 7 |
|---|---|---|---|---|---|---|
| • Last Name<br>• Full DOB<br>• Zip Code | • Last Name<br>• Full DOB<br>• Sex | • First Name<br>• Full DOB<br>• Sex | • State ID<br>• Full DOB<br>• Sex | • Last Name<br>• Month of Birth<br>• State<br>• Zip Code | • Last Name<br>• First Name<br>• State<br>• Sex | • Year of Birth<br>• Sex<br>• State<br>• Zip Code |
| **Score 1** | **Score 2** | **Score 3** | **Score 4** | **Score 5** | **Score 6** | **Score 7** |
| • First Name<br>• Middle Name<br>• Sex | • First Name<br>• Middle Name<br>• State<br>• Zip Code | • Middle Name<br>• Last Name<br>• State<br>• Zip Code | • First Name<br>• Middle Name<br>• Last Name<br>• Zip Code | • First Name<br>• Middle Name<br>• Year of Birth<br>• Day of Birth<br>• Sex | • Middle Name<br>• Month of Birth<br>• Year of Birth<br>• Day of Birth<br>• Zip Code | • First Name<br>• Middle Name<br>• Last Name<br>• Month of Birth<br>• Day of Birth |

## 2.2 Score Linked Pairs

Next, each linked pair 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 2.3), which allows the linkage algorithm to select final pairs to include in the matched file. The scoring process follows the following order:

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

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

- First Name or First Initial (when applicable)

- Middle Initial
- Last Name (Conditional on sex) or Last Initial (when applicable)
- Year-of-Birth
- Month-of-Birth
- Day-of-Birth
- Sex
- State-of-Residence
- Zip Code

### 2.2.1 Calculate M- and U- Probabilities

The **M-probability** – the probability that the identifiers from the records in question agree, given that the two records are a match – are computed separately within each individual block based on the scoring identifiers listed in table 1. Within the block, linked pairs with non-missing and matching (5 or more digits in agreement) SSN are used to calculate the M-probability, these links are assumed to represent the same individual. Only the variables outside of the blocking factor have an M-probability calculated. For example, among the assumed true positive pairs in block 2, if 99.4% agree on first name and 99.7% agree on state of residence, these percentages represent the M-probabilities for these identifiers.

First and last name identifiers have several additional comparison measures created for use in the calculation of M-probabilities:

- First/last initial – used in the scoring process when only an initial is present in the name field
- Jaro-Winkler Similarity Levels – this process is explained in greater detail in the section following the U-probability
- Last name is conditional on sex – it is common practice that women change their maiden name to their spouse's last name after marriage. This results in a lower agreement weight amongst the female population and should be taken into consideration.

The **U-probability** – the probability that the two values for an identifier from paired records agree given that they are NOT a match. With the exception of first and last names, these probabilities are calculated within each block, using records in which non-missing SSN are not matching (less than 5 digits in agreement).

Like the M-probabilities, only variables outside of the blocking factor have a U-probability calculated. The U-probabilities are calculated using the frequency of each value with in a variable in the full set of NHCS submission records. For example, the U-probability of a state with many patients would be around 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. Similar to the M-probabilities, first and last name are not calculated in the same manner as the rest of the identifiers. The calculation of the U-probabilities for first and last name are discussed in much greater detail in the following section.

### 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. Because of the many possible values for first and last name, it was impractical to compute U- probabilities specific to each blocking factor. Instead, we computed U-probabilities using all records in the NHCS submission file and a simple random sample of 1% of the CMS Medicare EDB 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, we randomly selected 100,000 names from the CMS Medicare EDB submission file 1% 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, and for each NHCS name, we tallied the number of the 100,000 randomly selected CMS Medicare EDB names that agreed at that level.[10,11]

### 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 calculated differently for each linked record pair, but follow a same general process:

- 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 (weight of zero)

First Name and Last Name weights are assigned using Jaro-Winkler similarity scores described in the above section. These scores range from 0 to 1, with 0 being a complete non-match and 1 being an exact match. The weighting algorithm assigns all scores below 0.85 a disagreement weight. All scores above 0.85 are assigned an agreement weight associated with the 85% level. If there is agreement at the 0.85 level then the pair is assessed at the 0.90 level. If the names disagree at this level, they are assigned a disagreement weight, although a much smaller value than the one assigned for failing the 0.85 check. If they agree, it is assigned an additional

[10] 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.
[11] 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.

agreement weight. This process continues two more times for the 0.95 and 1.0 thresholds, respectively.

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-specific probabilities allows the linkage algorithm (as described in Section 2.2) 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.

The logistic model was estimated on possible matched pairs that had a valid and formatted SSN on both the NHCS patient record and the CMS enrollment record. 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).[12] 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.

A logistic regression model is applied to each of the blocks in the block scheme, using full pair weight as the only variable in the regression model. Each of the resulting models produce an estimated predictive probability of match validity, calculated using the betas from the corresponding model, for records without a valid SSN/HICN.

# 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[13] using the logistic model.

The penultimate step is to assign a probability threshold that a pair is a valid match. The probability threshold for the linked 2014 NHCS – 2014/2015 CMS Medicare MBSF is set to 0.85. 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.

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

[12] 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.
[13] The probabilistic linkage match validity is estimated by logistic regression, value between zero and 1 (non-inclusive).

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.