The Linkage of the 2019 and 2020 National Hospital Care Survey (NHCS) to 2019–2020 and 2020–2021 Medicare Enrollment and Claims/Encounters Data: Linkage Methodology and Analytic Considerations

Data Release Date: March 16, 2026
Document Version Date: February 27, 2026

Division of Analysis and Epidemiology
National Center for Health Statistics
Centers for Disease Control and Prevention
datalinkage@cdc.gov

Suggested Citation: National Center for Health Statistics. Division of Analysis and Epidemiology. *The Linkage of the 2019 and 2020 National Hospital Care Survey to 2019–2020 and 2020–2021 Medicare Enrollment and Claims/Encounters Data: Linkage Methodology and Analytic Considerations*, March 2026. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/linked-data/nhcs/restricted-cms-medicare.html

Contents

List of Acronyms

AMA, American Medical Association
CCW, Chronic Conditions Warehouse
CMS, Center for Medicare & Medicaid Services
CPT-4, Current Procedural Terminology, 4th Edition
DHCS, Division of Health Care Statistics
DME, durable medical equipment
DMERC, durable medical equipment regional carrier
DOB, date of birth
DSH, disproportionate share
EDB, enrollment database
EHR, electronic health record
E-M, expectation-maximization
ERB, Ethics Review Board
ESRD, end-stage renal disease
FFS, fee-for-service
GME, graduate medical education
HCPCS, Healthcare Common Procedure Coding System
HHA, home health agency
HMO, health maintenance organization
HUD, Department of Housing and Urban Development
ICD-10-CM/PCS, International Classification of Diseases, 10th edition, Clinical Modification/Procedure Classification System
IME, indirect medical education
IP, inpatient
MA, Medicare Advantage
MAC, Medicare Administrative Contractor
MAO, Medicare Advantage Organization
MA-PD, Medicare Advantage Prescription Drug Plan
MBSF, Master Beneficiary Summary File
MedPAR, Medicare Provider Analysis and Review File
NCHS, National Center for Health Statistics
NDI, National Death Index
NHCS, National Hospital Care Survey
OP, outpatient
OTC, over-the-counter
PDE, prescription drug event
PDP, prescription drug plan
PII, personally identifiable information
RDC, Research Data Center
ResDAC, Research Data Assistance Center
SAF, standard analytic file

SNF, skilled nursing facility
SSN, Social Security number
VRDC, Virtual Research Data Center

1 Introduction

As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to provide statistical information that can be used to guide actions and policy to improve the health of the American people. As part of its ongoing efforts to fulfill this mission, NCHS conducts several population-based surveys and establishment surveys, including the National Hospital Care Survey (NHCS), https://www.cdc.gov/nchs/nhcs/index.html.

Through its data linkage program, NCHS has been able to expand the analytic utility of the data collected from NHCS by augmenting it with Medicare data collected by the Centers for Medicare & Medicaid Services (CMS). The linkage of NHCS patient data with CMS Medicare data creates new data resources that can support a wide range of public health surveillance and policy evaluation studies.

This report includes a brief overview of the linked data sources, the methods used for linkage, and analytic guidance. For more information or questions about the NHCS and Medicare linkage, please visit the data linkage website or contact the NCHS Data Linkage Program at datalinkage@cdc.gov.

2 Data Sources

2.1 National Hospital Care Survey (NHCS)

NHCS is an establishment survey that collects inpatient (IP) and emergency department (ED) episode-level data from sampled hospitals. It is one of the National Health Care Surveys, a family of surveys covering a wide spectrum of healthcare delivery settings including ambulatory, hospital, and post-acute and long-term care providers. The NHCS includes detailed information about hospital characteristics, patients' characteristics, and treatment. 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, health services utilization, and substance-involved ED visits.

From participating hospitals, NHCS collects data on all IP and ED 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 2019 and 2020 data collection hospitals were given the option of providing their data in the form of either Uniform Billing (UB)-04 administrative claims records or electronic health records (EHR). Additionally, data could be provided by third-party entities, like Vizient or (starting in 2020) the American College of Emergency Physicians. Participating hospitals were required to submit one type of data (e.g., UB-04 administrative claims or EHR, not both). For those hospitals submitting EHR data, this was submitted in the format of HL7 CDA® R2 Implementation Guide: National Health Care Surveys Release 1, DSTU Release 1.2 – US Realm (http://www.hl7.org/implement/standards/product_brief.cfm?product_id=385).

The 2019 NHCS collected data from a sample of 598 hospitals of which 82 provided linkage-eligible patient data and the 2020 NHCS collected data from a sample of 608 hospitals of which 106 provided linkage-eligible patient data. For participating hospitals, these data cover all hospital encounters to the inpatient and emergency department occurring throughout the calendar year. Even though NHCS is an establishment survey (i.e., hospitals are the sampling unit), it collects patient personally identifiable information (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 CMS Medicare

data. The linkage described here includes only IP and ED visits.

2.2 Centers for Medicare & Medicaid Services, Medicare Data

Medicare is the federal health insurance program for people age 65 or older, people under age 65 with qualifying disabilities, and people of all ages with end-stage renal disease (ESRD).

During 2019–2021, approximately 60% of persons enrolled in Medicare, known as Medicare beneficiaries, were enrolled in Original Medicare, also known as Medicare FFS, and 40% of beneficiaries received Medicare benefits through a Medicare Advantage (MA) plan, also known as Medicare Part C. [1]

Beginning in 2006, Medicare beneficiaries could elect optional prescription drug coverage, known as Medicare Part D. Part D coverage can be obtained through Medicare approved Part D private plans, known as Prescription Drug Plans (PDPs) or through Medicare Advantage Prescription Drug Plans (MA-PDs). Approximately 75% of Medicare beneficiaries are enrolled in a prescription drug plan.[1]

The CMS Medicare Data Files are comprised of Standard Analytic Files, or SAFs, which contain information on the enrollment status, health care utilization, and expenditures of Medicare-enrolled beneficiaries.

The SAFs for Medicare beneficiaries enrolled in FFS Medicare contain final action health care claims. A final action claim contains all payment adjustments between Medicare and providers and represents Medicare's final payment action for a given health care claim. Medicare FFS SAFs are organized by seven health care settings: inpatient (IP) hospital care, skilled nursing facility (SNF) stays, institutional outpatient (OP) care, practitioner/provider services (Carrier), home health agency (HHA), durable medical equipment (DME), and hospice care.

The SAFs for MA-enrolled beneficiaries contain all health care encounter records. MA SAFs are organized by six health care settings: IP, SNFs, OP, Carrier, HHA, and DME. Hospice care services provided to Medicare beneficiaries enrolled in MA are paid under Medicare FFS rather than as part of the managed care plan.

The Medicare Part D Prescription Drug Event (PDE) File contains a summary of prescription drug costs and payment data used by CMS to administer benefits for all Medicare Part D enrollees, including beneficiaries enrolled in both Medicare PDPs and MA-PDs.

For a more detailed description of the information included in each of the Medicare Data Files, please see Appendix I: Descriptions of Medicare Data Files.

3 Linkage Methodology

3.1 Linkage Eligibility Determination

The linkage of 2019 and 2020 NHCS patient records to Medicare administrative data was conducted under an interagency agreement between NCHS and CMS. The linkage was performed in the CMS Virtual Research Data Center (VRDC). Approval for the linkage was provided by NCHS's Research Ethics

[1] CMS Medicare Enrollment Dashboard. https://data.cms.gov/tools/medicare-enrollment-dashboard

Review Board (ERB).[2]

Linkage was attempted only for NHCS patient records that had at least two of the following three identifiers present:

- valid SSN [3]
- valid date of birth (month, day, and year)[4]
- valid name (first, middle initial, and last)[5]

For example, if the PII on the NHCS 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.

The variable ELIGSTAT, included on the Match Status file, provides the linkage eligibility status for each NHCS patient record: ELIGSTAT values include 0 (ineligible) or 1 (eligible). It should be noted that linkage eligibility is distinct from program eligibility, which defines whether a person meets federal and state-specific eligibility criteria for a specific government-administered or-funded program.

3.2 Overview of Linkage

This section outlines the steps used to link the 2019 and 2020 NHCS data to the CMS Medicare Enrollment Database (EDB). For more detailed information on linkage methodology see Appendix II: Detailed Description of Linkage Methodology.

Linkage-eligible NHCS patient records were linked to records in the CMS EDB 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.

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

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

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

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

The NHCS patient records and the CMS EDB records were linked using both deterministic and probabilistic approaches. For the probabilistic approach, scoring was conducted according to the Fellegi-Sunter method.[6] Following this, a selection process was implemented with the goal of selecting pairs that represented the same individual between the two data sources. The following steps were implemented:

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

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

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

Upon completion of the linkage, a file containing the encrypted NCHS identification number and Medicare beneficiary identification number for successfully matched survey participants was provided to CMS VRDC staff. CMS extracted data records from its SAFs for successful matched NCHS survey participants and encrypted data files were shipped to NCHS, where additional quality control checks were performed.

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

3.3 Linkage Rates

Table 1 presents the total number of 2019 and 2020 NHCS patients by age group and sex, the number who were eligible for linkage, the number who were linked to CMS Medicare, and the unweighted percentage of all patients and those eligible for linkage who were linked to Medicare data by age and sex. Medicare has age-based entitlement at age 65. Therefore, the linkage rates for each survey were examined overall and by two age groups – 0–64 years and 65 years and older. Age was defined as the patient's age at their final IP or ED encounter date.

Table 1. Linked NHCS – CMS Medicare Records: Sample Sizes and Percent Linked, by Age and Sex

	Sample Size Total Sample	Sample Size Eligible for Linkage[2]	Sample Size Linked to Medicare[3]	Percent Linked Total Sample[4]	Percent Linked Eligible Sample[5]
2019 NHCS					
Age[1]					
0 - 64	2,286,950	2,171,492	159,661	6.98	7.35
65 and older	602,283	573,250	556,816	92.45	97.13
Not calculated	1,885,931	95	1	0.00	1.05
Total	4,775,164	2,744,837	716,478	15.00	26.10
Sex					
Male	1,307,165	1,244,490	325,651	24.91	26.17
Female	1,582,306	1,499,828	390,757	24.70	26.05
Missing	1,885,693	519	70	0.00	13.49
Total	4,775,164	2,744,837	716,478	15.00	26.10
2020 NHCS					
Age[1]					
0 - 64	2,517,479	2,415,012	190,672	7.57	7.90
65 and older	727,233	702,114	682,145	93.80	97.16
Not calculated	2,223,674	334	7	0.00	2.10
Total	5,468,386	3,117,460	872,824	15.96	28.00
Sex					
Male	1,497,364	1,425,121	409,390	27.34	28.73
Female	1,780,942	1,691,749	463,305	26.01	27.39
Missing	2,190,080	590	129	0.01	21.86
Total	5,468,386	3,117,460	872,824	15.96	28.00

NOTES: Data are presented at patient level.

[1] Age is as of final IP or ED encounter date. Age is calculated by subtracting patient date of birth (DOB) from the final encounter date. When more than one DOB was present, the minimum of the non-missing DOB was selected.

[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 linkage-eligible patients who linked to Medicare administrative records at any time during the linkage interval (2019 NHCS: 2019–2020 Medicare data, 2020 NHCS: 2020–2021 Medicare data).

[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 2019 and 2020 NHCS and CMS Medicare records. It is not an exhaustive list of analytic issues that researchers may encounter while using the linked NHCS-CMS Medicare data. Questions about analytic issues can be reported at (datalinkage@cdc.gov). Users of the linked NHCS - Medicare data files are encouraged to visit the ResDAC website (http://www.resdac.org) for additional information on Medicare data and their analytic considerations.

4.1 Analytic Considerations for Linked NHCS Data

4.1.1 NHCS Restricted-Use Files (RUF)

The 2019 and 2020 NHCS restricted-use survey data are made available for research use through the NCHS RDC network. For more information about obtaining access to NHCS RUFs see Section 5.1. The NHCS RUFs are organized as relational data tables organized by Facility, Patient, Encounter, Conditions, Services, Current Procedural Terminology (CPT)/Healthcare Common Procedure Coding System (HCPCS) Flag, Revenue Code Flag, and Encounter Weights. For more information about the specific variables and the observational unit for each data table please see the NHCS data documentation available at: https://www.cdc.gov/rdc/data/b1/NHCS-RDC-Data-Dictionary.pdf.

4.1.2 NHCS Hospital Eligibility and Sampling

Eligible hospitals for NHCS are non-institutional, non-federal hospitals with six or more staffed IP beds. There were 6,906 hospitals which met these criteria as of 2020 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 2017, the sample and frame files were updated to include newly constructed hospitals from a new source file. Updates to the NHCS sample and frame occur every three years. Due to the addition of newly sampled birth hospitals, the sample increased to 598 hospitals in 2019 and 608 hospitals in 2020. Moreover, of the 598 sampled hospitals in 2019, 82 hospitals were eligible for linkage and of the 608 sampled hospitals in 2020, 106 hospitals were eligible for linkage (note: the linkage-eligible number excludes hospitals that provided records covering less than 6 months of the analysis period). In 2019, of the 82 linkage-eligible hospitals, 82 hospitals sent IP data, and 76 hospitals sent ED data. Of the 106 linkage-eligible hospitals in 2020, 104 hospitals sent IP data, and 100 hospitals sent ED data. Although NHCS collected outpatient data from 2013-2016, outpatient data are no longer being collected.

4.1.3 2020 NHCS Encounter Weights

While national estimates of hospital encounters in the IP and ED are not available for NHCS 2019 due to low response rates, the 2020 NHCS can produce nationally representative estimates. The Division of Healthcare Statistics (DHCS) at NCHS has produced encounter level weights that can be used for national estimates of hospital encounters in the IP and ED. For information regarding producing weighted estimates with 2020 NHCS data the NHCS data documentation available at: https://www.cdc.gov/nchs/data/nhcs/2020-NHCS-PUF-Tech-Doc-508.pdf

The linkage between the NHCS data and CMS Medicare data was conducted at a patient level using patient identifiers collected from patient encounter records. The patient identifiers collected from the encounter records were used to link Medicare administrative records for NHCS patients. Although DHCS has developed encounter level weights for use with the NHCS 2020 encounter data, patient level

weights for use with linked Medicare data are not available.

4.1.4 NHCS Patient Identification Number

Each patient in the NHCS is assigned a unique identification number, PATIENT_ID. PATIENT_ID does not contain any identifiable information about the patient and is intended to be unique for each individual receiving IP or ED services at a participating hospital. However, 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.[7]

4.2 Analytic Considerations for Linked Medicare Data Files

Records for 2019 NHCS have been linked to 2019–2020 records, and records for 2020 NHCS have been linked to 2020–2021 records, from the following CMS Medicare Data Files:

- Master Beneficiary Summary File (MBSF)
    - Base (Medicare Parts A/B/C/D) Segment
    - Cost & Utilization Segment
    - Chronic Conditions Segments
- Part D Prescription Drug Event (PDE)
- Medicare Provider Analysis and Review File (MedPAR)
- Fee-for-Service (Claim files)
    - Inpatient (IP)
    - Skilled Nursing Facility (SNF)
    - Professional (Carrier)
    - Outpatient (OP)
    - Durable Medical Equipment (DME)
    - Home Health Agency (HHA)
    - Hospice
- Medicare Advantage (Encounter Files)
    - Inpatient (IP)
    - Skilled Nursing Facility (SNF)
    - Professional (Carrier)
    - Outpatient (OP)
    - Durable Medical Equipment (DME)
    - Home Health Agency (HHA)

Important Note: All RDC applications to analyze linked NHCS-CMS Medicare data should include requests to analyze the MBSF Base Segment File for the same calendar year(s) as the Medicare health care claims, encounter, or prescription drug data to allow researchers to determine the correct study denominators for the various Medicare programs (Medicare Parts A, B, C, and D). The MBSF includes critically important information on Medicare program entitlement and enrollment and should always be used in conjunction with other Medicare data files to identify Medicare beneficiaries eligible for service utilization within each program.

[7] For more information of Patient_ID generation, see Technical Notes on page 14: https://www.cdc.gov/nchs/data/nhsr/nhsr097.pdf.

More detailed descriptions of the linked Medicare data files are provided in Appendix I. The following sections address potential analytic considerations specific to each of the linked Medicare data files.

4.2.1 Match Status File

The Match Status File can be used to identify which of the NHCS patients were eligible for linkage and linked to a Medicare record. This file contains one record for each unique NHCS patient ID and contains the variables ELIGSTAT, PROBVALID, and MEDICARE_MATCH_STATUS.

The variable ELIGSTAT should be used to determine linkage eligibility (Section 3.1). NHCS patient IDs with an ELIGSTAT value of 1 were considered eligible for linkage to CMS Medicare records.

This file also contains information on the estimated probability of match validity (PROBVALID). An estimated probability of match validity was computed for each candidate pair and compared against a probability cut-off value to determine which pairs were links (an inferred match). For additional discussion on how PROBVALID was estimated, see Appendix II Section 3.3 and 3.4. NCHS used a probability cut-off value which minimized the total estimated counts of Type I error (false positive links – identified as enrolled in Medicare but actually are not) and Type II error (false negative links – identified as not enrolled in Medicare but actually are).

In the 2019 and 2020 NHCS – CMS Medicare linkages, NCHS used a probability cut-off value of 0.85 to determine final match status. Candidate pairs with a PROBVALID that exceeded the probability cut-off value (i.e., PROBVALID>0.85) were deemed a link. For additional discussion on probability cut-off value determination and record selection, please see Appendix II Section 4. For some analyses, it may be desirable to reduce the Type I error. To do this, researchers should increase the probability cut-off value to a value closer to 1.0. Researchers wishing to increase the probability cut-off value should request PROBVALID in their RDC proposal. Note, the probability cut-off value cannot be decreased from 0.85 as pairs estimated with lower match probability are not made available to researchers.

The MEDICARE_MATCH_STATUS variable can be used to identify which of the NHCS patients were linked to Medicare administrative records at any time during the Medicare linkage period. When equal to one, MEDICARE_MATCH_STATUS indicates that a NHCS patient was matched to any Medicare administrative records during the linkage period, either 2019–2020 for NHCS 2019 or 2020–2021 for NHCS 2020.

4.2.2 Analytic Considerations Specific to the Master Beneficiary Summary File (MBSF)

The MBSF provides data on linked NHCS-Medicare beneficiaries enrolled in a Medicare program at some point during the MBSF reference year. Reference year refers specifically to the calendar year accounted for in the linked MBSF. For example, the linked 2019 NHCS and 2019 MBSF will contain information for Medicare enrollment and summary health care utilization occurring in 2019.

4.2.2.1 MBSF File Year Indicator

The MBSF reference year can be found in the variables BENE_ENROLLMT_REF_YR and FILE_YEAR4. **All research proposals should include one of these variables.** Please note that linked records for all years of Medicare enrollment are appended into a single file. Many beneficiaries are enrolled in Medicare during multiple years of the linkage period and thus appear multiple times in the file.

4.2.2.2 MBSF Base Segment File (Medicare Parts A/B/C/D)

Creating Medicare Study Denominators

> **Note:** To properly construct linked NHCS-CMS Medicare study populations researchers must request and use the MBSF to determine the correct study denominators for each Medicare program (Medicare Parts A, B, C, and D). The MBSF includes critically important information on Medicare program entitlement and enrollment.

The linked MBSF Base (A/B/C/D) segment includes essential information to create study denominators. Monthly enrollment variables indicate when a given linked NHCS patient was enrolled in specific Medicare programs during the year. These indicators can be used to determine which beneficiaries were eligible to receive covered health services in each Medicare program. For example, beneficiaries who are not enrolled in Medicare Part B will not have health care claims for services paid under it – including physician visits, OP procedures, HHA services, or DME. Beneficiaries enrolled in MA or Medicare Part C will not have health care claims data but will instead have health care encounter records reported by their Medicare Advantage Organization (MAO).

Indicators for Part A and B entitlement for each month of the calendar year are provided in the variables MDCR_STATUS_CODE_01 - MDCR_STATUS_CODE_12. MA enrollment monthly indicators are found in HMO_IND_01 - HMO_IND_12. Part D has no monthly enrollment indicator variable, but for any value of PTD_CNTRCT_ID_01 - PTD_CNTRCT_ID_12 that is N, 0, or null/missing for that month, the beneficiary did not have Part D coverage for that month. There may be instances where a linked NHCS patient is enrolled in Medicare FFS or MA but no FFS claims or Medicare encounter records are available, because it is possible to be enrolled in Medicare but not utilize Medicare services during the coverage period for a given calendar year.

For more information on how to create an analytic sample that excludes Medicare beneficiaries enrolled in a MA plan, refer to a document written by ResDAC: https://www.resdac.org/articles/identifying-medicare-managed-care-beneficiaries-master-beneficiary-summary-or-denominator. Additional analytic considerations specific to analyzing data for MA enrollees are provided in Section 4.2.4.

Medicare Entitlement

The linked MBSF Base (A/B/C/D) segment also includes three variables indicating Medicare entitlement: original reason for entitlement, current reason for entitlement, and Medicare status code.

- A beneficiary's *original reason* for Medicare entitlement is found in the variable ENTLMT_RSN_ORIG. Knowing a beneficiary's original reason for entitlement can be useful for identifying which aged beneficiaries were formerly entitled (i.e., prior to age 65) to Medicare due to a qualifying disability. Possible values include: Old Age and Survivors Insurance (OASI), Disability Insurance Benefits (DIB) and 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.

Race and Ethnicity

The linked MBSF Base (A/B/C/D) Segment provides two race and ethnicity variables: BENE_RACE_CD, which is the variable reported in the CMS administrative claims data system, and RTI_RACE_CD, which contains race and ethnicity codes imputed through the use of an algorithm developed by the Research Triangle Institute (RTI) to improve the accuracy of race and ethnicity data included 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.

4.2.2.3 MBSF Cost and Use Segment

The linked MBSF Cost and Use segment includes one record for each beneficiary enrolled in FFS Medicare in the calendar year of the file. This record includes summary utilization and total annual payment for FFS Medicare-covered services including hospitalizations and physician visits. Additional information about the variables included in the linked NHCS MBSF Cost and Use segment is available at https://resdac.org/cms-data/files/mbsf-cost-and-use.

4.2.2.4 MBSF Chronic Conditions Segments

The CMS Medicare MBSF Chronic Conditions segments include variables indicating whether each Medicare FFS-enrolled beneficiary has claims indicating the presence of multiple specific chronic conditions. The linked data includes two versions of the Chronic Conditions categories: the 27 CCW Chronic Conditions file and the 30 CCW Chronic Conditions file, which uses enhanced algorithms. CMS provides additional information about the methodology used to assign chronic condition flags to Medicare beneficiaries on their website (https://www2.ccwdata.org/web/guest/condition-categories-chronic) and in the Chronic Conditions File Enhancement White Paper (available at https://www2.ccwdata.org/documents/10280/19002256/ccw-condition-categories-impact-of-transition-from-27-to-30.pdf).

Please note: According to CMS documentation, 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).

4.2.2.5 MBSF Other Chronic or Potentially Disabling Conditions

The CMS Medicare MBSF Other Chronic or Potentially Disabling Conditions segment include variables indicating whether each Medicare FFS-enrolled beneficiary has claims indicating the presence of multiple specific conditions not included in the original list of 27 conditions. CCW provides additional information about the methodology used to assign these other conditions flags to Medicare beneficiaries on their website (https://www2.ccwdata.org/web/guest/condition-categories-other).

Please note: According to CMS documentation, 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).

4.2.3 Analytic Considerations Specific to Medicare Fee-for-Service Claims Files

The Medicare FFS Claims Files contain information from claims for reimbursement for health care services provided to Medicare beneficiaries enrolled in FFS or Original Medicare (Medicare Part A and/or Part B). Claims submitted for reimbursement from institutional providers (Medicare Part A) include IP, OP, SNFs, HHAs, and Hospice Services and are paid under the rules published for the prospective payment systems established for institutional providers. Claims submitted for reimbursement for non-institutional providers including professional providers (e.g. doctors, physician assistants) and providers of DME (Medicare Part B) are paid according to published fee schedules.

The data provided on the linked NHCS-Medicare FFS Files represent the final adjudication of the Medicare payment amount of each health care claim. However, the final Medicare payment amount may not represent the full cost of health care services provided to Medicare beneficiaries. Medicare beneficiaries can be subject to cost sharing requirements (i.e. deductibles and coinsurance) for Medicare covered health care services. It is not possible to determine whether the beneficiary paid the cost-sharing amount "out-of-pocket" or whether the cost-sharing amounts are paid by a third party, such as a Medi-gap policy. Therefore, the total amount spent for a given health care service may not be captured by relying on the Medicare FFS claims payment data alone. CMS has published additional guidance to assist with analysis of Medicare FFS claims data in the CCW Medicare Data User Guide.

4.2.3.1 Carrier File

The claims on the FFS Carrier File are processed by private carriers working under contract to CMS. Each carrier claim includes a Healthcare Common Procedure Coding System (HCPCS) code to describe the nature of the billed service. The HCPCS are composed primarily of Level I HCPCS or Current Procedural Terminology (CPT-4) codes developed by the American Medical Association (AMA), with additional CMS specific codes called Level II HCPCS. Level II HCPCS are used to identify products, supplies, and services that are not included in AMA's CPT codes. These may include ambulance services, DME, prosthetics, and orthotics. Each HCPCS code on the carrier claim must be accompanied by a diagnosis code based on the International Classification of Diseases, Tenth Revision, Clinical Modification / Procedure Coding System (ICD–10–CM/PCS), providing a reason for the service. In addition, each record includes the date of service and reimbursement amount.

Providers, such as physicians, can bill for services provided in the office, hospital, or other sites. The Line Place of Service Code (LINE_PLACE_OF_SRVC_CD) indicates where the service was provided, but it is not required for payment purposes.

The FFS Carrier File contains DME claims processed by payment contractors who also process physician claims. The DME line items included on the FFS Carrier File can be identified by Claim Type Code (NCH_CLM_TYPE_CD) equal to 72. DME claims processed through DME regional carriers are found on the FFS DME Files, not on the Carrier File. DME claims on the Carrier File are for separate services. For additional information on DME regional carrier claims, see the DME File description in Section 4.2.3.2.

The Carrier File has two pairs of date fields. Claim From Date (CLM_FROM_DT) and Claim Through Date (CLM_THRU_DT) generally cover a period of service (but not always a single date of service), while Line First Expense Date (LINE_1ST_EXPNS_DT) and Line Last Expense Date (LINE_LAST_EXPNS_DT) represent the specific day of the provided service.

For every billed procedure (using an HCPCS code), a corresponding ICD–10–CM diagnosis code (LINE_ICD_DGNS_CD) should appear providing the reason for the billed service.

4.2.3.2 Durable Medical Equipment (DME) File

Durable medical equipment or DME can be billed through either a) the carriers who also process physician claims, or b) DME Regional Carriers (DMERCs), who process only DME claims

DME claims processed by suppliers who also process physician claims are included only on the FFS Carrier File. These claims can be identified by Claim Type Code (NCH_CLM_TYPE_CD) equal to 72 on the Carrier File. DME claims processed by regional carriers are included only on the FFS DME File.

4.2.3.3 Hospice File

Physician claims included in the Hospice File are for services provided by physicians employed or receiving payment from the hospice facility. All hospice claims are processed as Medicare claims regardless of whether the beneficiary is enrolled in an FFS or MA plan.

4.2.3.4 Outpatient (OP) File

Same-day surgeries performed in a hospital are included in the FFS OP File. However, claims for surgeries performed in freestanding surgical centers appear in the FFS Carrier File, not in the FFS OP File.

4.2.3.5 Inpatient (IP) File

Each record on this file represents a health care claim submitted for payment by inpatient hospital providers for reimbursement of facility costs incurred during the provision of inpatient care. Multiple claims records may be submitted for one hospital stay. Researchers interested in analyzing summarized information for inpatient stays rather than individual inpatient claims may wish to use the MedPAR file (described in Section 4.2.3.7) which summarizes individual inpatient claims at the stay level. Researchers interested in analyzing inpatient data across the FFS and MA programs should use the FFS and MA Inpatient Files as there is currently no MedPAR type data file created to summarize Inpatient encounters at the stay level for the MA program.

Observation care services that result in an inpatient admission within 3 days of the start of the observation period will be included in the Inpatient File and can be identified with a revenue center code 0762. Observation care provided in the Inpatient setting, but which does not result in an inpatient admission within 3 days of the start of the observation period are included on the FFS OP File.

4.2.3.6 Skilled Nursing Facility (SNF) File

Each claim record on this file represents a health care claim submitted for payment by a SNF for reimbursement of the provision of skilled nursing care. Multiple claims records may be submitted for one SNF stay. Medicare billing frequency guidance for SNFs requires SNFs to submit claims at least monthly. Researchers interested in analyzing claims information summarized at the stay level may wish to use the MedPAR file which summarizes individual SNF claims at the stay level (see Section 4.2.3.7). Researchers interested in analyzing SNF data across the FFS and MA programs should use the FFS and MA SNF Files as there is currently no MedPAR type data file created to summarize SNF encounters at the stay level for the MA program.

4.2.3.7 Medicare Provider Analysis and Review (MedPAR) File

The MedPAR file creates a single summarized record for each hospital or SNF stay, containing information on ICD-10-CM/PCS codes, admission, discharge, and procedure dates from the individual IP and SNF final action claims. Information regarding charges for IP or SNF services are more highly aggregated in MedPAR than those provided in the Inpatient and SNF Claims Files. Each MedPAR record

may represent one IP or SNF claim or multiple claims, depending on the length of a beneficiary's stay and the amount of services billed throughout the stay. Researchers interested in the more granular detail of individual IP or SNF claims should use the FFS IP or SNF Claims Files for their analyses.

The MedPAR file includes all hospitalizations that had a discharge date during the calendar year and all SNF stays with an admission date during the calendar year. Hospital stays starting in one calendar year and continuing past the end of the calendar year are not included in the MedPAR file until the year of discharge. To determine if a record is for a long- or short-stay hospitalization, use the short stay/long stay/SNF indicator variable SS_LS_SNF_IND_CD which is coded 'S' for short stay or 'L' for long stay.

The MedPAR files may include "information only" claims for MA-enrolled beneficiaries that are submitted by IP and SNF facilities for calculation of disproportionate share (DSH), indirect medical education (IME) and graduate medical education (GME) payments. Note that CMS advises removing MA-covered claims from health care utilization analyses based on MedPAR data. For more information on removing information only claims from the MedPAR file see https://www.resdac.org/articles/identifying-medicare-managed-care-beneficiaries-master-beneficiary-summary-or-denominator. The CMS FFS IP and SNF Claims Files do not include "information only" claims.

All individual IP and SNF encounter records submitted by the MAOs are available for analysis on the linked IP and SNF Encounter Data Files.

4.2.4 Analytic Considerations Specific to Medicare Advantage (MA) Encounter Files

MA encounter data reflect services provided to Medicare beneficiaries enrolled in MA plans, also known as Medicare Part C. Unlike FFS claims, CMS does not use MA encounter data as the basis for payments to providers of health care services. Rather, CMS pays the MAOs a capitated payment amount per enrolled beneficiary.

There are 2 types of encounter data records that MAOs submit to CMS, Encounter Data Records and Chart Review Records.

Encounter data records capture information on health care services provided to MA-enrolled beneficiaries. MA encounter records differ from FFS claims because: 1) they are reported to CMS by MAOs rather than directly from the provider of health care services, 2) multiple encounter records may be reported for the same health care service, 3) NCHS_ENC_JOIN_KEY should be used to match together claims between the base and line/revenue claims files, 4) some encounter records contain service codes that are not used in FFS Medicare and 5) certain information on an encounter record may not always be fully populated if the information is not required for MAO payment purposes.

Chart review records are a type of MA encounter data record used by MAOs to add or remove diagnoses that they identify through medical record reviews. Chart review records can be submitted for any health care service type, and there is no limitation on the number of chart review records that a MAO may submit. MAOs have the option of submitting linked chart reviews which are linked to the original encounter data record or chart review record through the claim control number (i.e. NCHS_CLM_CNTL_NUM will be equal to NCHS_CLM_ORIG_CNTL_NUM of an original encounter or chart review record). Linked chart review records can be used to add or delete diagnoses previously reported or can be used to void a previously reported encounter record. Unlinked chart review records are not linked to an original encounter or chart review record. Unlinked chart review records can only be used

to add diagnoses. Chart review records can be identified by the variable Chart Review Switch (CLM_CHRT_RVW_SW).

CMS has published additional guidance to assist with analysis of Medicare encounter claims data in the CCW Medicare Encounter Data User Guide.

4.2.5 Analytic Considerations Specific to Medicare Part D Prescription Drug Event (PDE) File

Medicare Prescription Drug coverage or Medicare Part D is provided by PDPs, which offer only prescription drug coverage, or through MA-PD plans, which offer prescription drug coverage that is integrated with the health care coverage provided to Medicare beneficiaries under Medicare Advantage plans. The PDE file includes prescription drug event data for beneficiaries enrolled in either PDPs or MA-PDs. The PDE file contains summary extracts submitted to CMS by Medicare Part D PDP providers.

CMS has published additional guidance to assist with analysis of Medicare prescription drug data in the CCW Medicare Part D Data User Guide.

5 Access to Data Files

5.1 Access to the Restricted-Use Linked NHCS-CMS Medicare Data Files

To ensure confidentiality of data, NCHS provides safeguards including the removal of all personal identifiers from analytic files. Additionally, the linked data files are only accessible through the NCHS Research Data Center (RDC) network for approved research projects. Researchers who wish to access the restricted-use 2019 or 2020 NHCS survey files and the linked 2019 or 2020 NHCS-CMS Medicare data files must submit a research proposal application to the NCHS RDC. The RDC staff will review all submitted proposals to determine if the proposed project is feasible and to identify any potential disclosure risks. More information regarding the NCHS RDC network and the RDC proposal application process are available from: https://www.cdc.gov/rdc/.

5.2 Merging the NHCS Analytic Files with the Linked NHCS-CMS Medicare Files

The linkage between the 2019 and 2020 NHCS data and CMS Medicare data was conducted at a patient level using patient-level identifiers. The shared variable, PATIENT_ID, will be used by the RDC to merge the linked 2019 or 2020 NHCS – CMS Medicare files with the restricted-use 2019 or 2020 NHCS data. Analysts should request all variables of interest from the NHCS restricted-use data files and the linked NHCS – CMS Medicare files in their RDC proposal.

5.3 Additional Related Data Sources

In addition to the linked NHCS-CMS Medicare data, researchers may also request variables from the linked 2019 NHCS–2019-2020 NDI data file and the linked 2020 NHCS–2020-2021 NDI data file if mortality is an outcome of interest (https://www.cdc.gov/nchs/linked-data/nhcs/restricted-ndi.html). The linked mortality file includes the NHCS Patient identification number, date of birth, date of death, and cause of death information for linked decedents. To integrate the NHCS linked mortality files with the linked NHCS-CMS Medicare data files, joins (merges) are made on the common identification number, PATIENT_ID.

Researchers interested in studies focused on the relationship between housing and health may also request variables from the linked 2019 NHCS-2018–2020 HUD administrative data or the linked 2020 NHCS-2019–2021 HUD administrative data (https://www.cdc.gov/nchs/linked-data/nhcs/restricted-hud.html). The linked HUD administrative data files include variables pertaining to the recipient's participation in Housing Choice Voucher (HCV), Public Housing (PH), and/or Multifamily (MF) programs. To integrate the linked NHCS-HUD administrative data files with the linked NHCS-CMS Medicare data files, joins (merges) are made on the common identification number, PATIENT_ID.

Appendix I Descriptions of Medicare Data Files

This appendix contains a brief description of the Medicare data files. Additional information may also be found at https://resdac.org/file-availability.

1 Master Beneficiary Summary File (MBSF)

The MBSF is an annual file containing demographic and enrollment information about beneficiaries enrolled in Medicare during each calendar year. The MBSF consists of three segments. The **Base (A/B/C/D) segment** includes beneficiary characteristics, monthly entitlement indicators, reasons for entitlement (initial and current), and monthly Medicare program enrollment indicators. The **Cost and Use segment** includes summarized information about the service utilization and Medicare payment information for Medicare beneficiaries enrolled in Medicare FFS by type of claim, including summary information on prescription drugs. The **Chronic Conditions segments** include variables that indicate a Medicare FFS-enrolled beneficiary has received a service or treatment for selected chronic health conditions.

2 Standard Analytic Files (SAFs)

The SAFs for Medicare beneficiaries enrolled in FFS Medicare contain final action health care claims submitted for payment by both institutional and non-institutional health care providers. A final action claim contains all payment adjustments between Medicare and providers and represents Medicare's final payment action for a given health care claim. Medicare FFS SAFs are organized by seven health care settings: IP, SNF, OP, Carrier, HHA, DME, and Hospice care.

The SAFs for MA-enrolled beneficiaries contain all health care encounter records submitted by MAOs for the given calendar year for each enrolled Medicare beneficiary. MA SAFs are organized by six health care settings: IP, SNF, OP, Carrier, HHA, and DME. Hospice care services provided to Medicare beneficiaries enrolled in MA are paid under Medicare FFS rather than as part of the managed care plan.

The data for the IP, SNF, OP, HHA, and Hospice files were all provided in a similar format. Each of the files are divided into seven segments: 1) a base claim segments including demographic information, diagnosis codes, procedures codes, and dates of service; 2) a condition segment, identifying the claim-related condition; 3) an occurrence code segment, identifying a notable claim-related event and date that may affect processing of payment by CMS; 4) a span code segment, identifying a notable claim-related event and time period that may affect payment processing; 5) a value code segment including the billing and reimbursement amounts associated with a claim; 6) a revenue code segment identifying the cost center or division/unit within a hospital in which a charge is billed; and 7) a demonstration code segment identifying claims processed as part of a CMS demonstration project. Each segment is available as a separate file, but can be combined using the unique claim identification number (NCHS_CLM_ID) or encounter join key (NCHS_ENC_JOIN_KEY), Medicare reference year (FILE_YEAR4) and unique patient identifier (PATIENT_ID).

The Carrier and DME files share similar formats. Each file consists of 1) a base claims segment, containing demographic information and diagnosis codes as well as billing and payment amounts associated with a non-institutionalized claim; 2) a line items segment that includes the specific billing and payment amounts for each line item included within the base claim; and 3) a demonstration code segment. The base claim, line item, and demonstration code segments are available as separate files but

can be combined using the unique claim identification number (NCHS_CLM_ID) or encounter join key (NCHS_ENC_JOIN_KEY), Medicare reference year (FILE_YEAR4) and unique patient identifier (PATIENT_ID).

2.1 Inpatient (IP) Files

2.1.1 Fee-for-Service Inpatient File

The FFS IP File contains Medicare Part A final action claims from IP facilities. The FFS IP File contains data fields for ICD-10-CM/PCS codes, revenue center codes, dates of service, and payment information. Each record on this file contains the information from one health care claim. Episodes of care may encompass more than one health care claim.

2.1.2 Encounter Inpatient File

The Encounter IP File contains health care encounters reported to CMS by MAOs in a format similar to the FFS IP claims, but encounter records do not include payment information. Additionally, chart review records, which allow MAOs to add or remove diagnoses from initially reported on values, are included on this file. The Encounter IP File contains encounter data submitted for the same types of institutional providers as those reported on the FFS IP File and may include encounter records reported for additional IP services provided by MA plans not covered by FFS Medicare. Episodes of care may encompass more than one health care encounter.

2.2 Skilled Nursing Facility (SNF) Files

2.2.1 Fee-for-Service SNF File

The FFS SNF File contains Medicare Part A final action claims from SNFs. The FFS SNF File contains data fields for ICD-10-CM/PCS codes, revenue center codes, dates of service, and payment information. Each record on this file contains the information from one health care claim. Episodes of care may encompass more than one health care claim. Skilled nursing care is the only level of nursing home care that is covered by the Medicare program.

2.2.2 Encounter SNF File

The Encounter SNF File contains health care encounters reported to CMS by MAOs in a format similar to the FFS SNF claims, but encounter records do not include payment information. Additionally, chart review records are included on this file and are a special type of MA encounter data that allows MAOs to add or remove diagnoses initially reported on encounter data records. The Encounter SNF File contains encounter data submitted for the same types of institutional providers as those reported on the FFS SNF File and may include encounter records reported for additional skilled nursing services provided by MA plans not covered by FFS Medicare. Episodes of care may encompass more than one health care encounter.

2.3 Carrier Files

2.3.1 Fee-for-Service Carrier File

The FFS Carrier File contains Medicare Part B final action claims data submitted by professional providers, including physicians, physician assistants, clinical social workers, and nurse practitioners. The data are largely made up of physician claim records but may also include claims for certain DME (see

Section 4.2.3.1) and claim records from certain organizational providers, such as independent clinical laboratories, ambulance providers, and free-standing ambulatory surgical centers. FFS Carrier claims include for ICD-10-CM/PCS codes, dates of service, and payment information. Each record on this file contains the information from one provider-submitted health care claim. Episodes of care may encompass more than one health care claim.

2.3.2 Encounter Carrier File

The Encounter Carrier File contains health care encounters reported to CMS by MAOs in a format similar to the FFS provider claims, but encounter records do not include payment information. Additionally, chart review records are included on this file and are a special type of MA encounter data that allows MAOs to add or remove diagnoses initially reported on encounter data records. The Encounter Carrier File contains encounter data submitted for the same types of providers as those reported on the FFS Carrier File and may include encounter records reported for additional services provided by MA plans not covered by FFS Medicare (such as dental, hearing or vision services). Episodes of care may encompass more than one health care encounter.

2.4 Outpatient (OP) Files

2.4.1 Fee-for-Service Outpatient File

The FFS OP File contains Medicare Part A final action claims from OP providers including: hospital OPDs, rural health clinics, renal dialysis facilities, OP rehabilitation facilities, comprehensive OP rehabilitation facilities, Federally Qualified Health Centers and community mental health centers. The FFS OP File contains data fields for ICD-10-CM/PCS codes, revenue center codes, dates of service, and payment information. Each record on this file contains the information from one health care claim. Episodes of care may encompass more than one health care claim.

2.4.2 Encounter Outpatient File

The Encounter OP File contains health care encounters reported to CMS by MAOs in a format similar to the FFS OP claims, but encounter records do not include payment information. Additionally, chart review records are also included on this file and are a special type of MA encounter data that allows MAOs to add or remove diagnoses initially reported on encounter data records. The Encounter OP File contains encounter data submitted for the same types of providers as those reported on the FFS OP File and may include encounter records reported for additional services provided by MA plans not covered by FFS Medicare (such as dental, hearing or vision services). Episodes of care may encompass more than one health care encounter.

2.5 Durable Medicare Equipment (DME) Files

2.5.1 Fee-for-Service DME File

The FFS DME File contains Medicare Part B final action claims data submitted by DME suppliers to a DME Medicare Administrative Contractor (MAC). Information in the FFS DME file includes for ICD-10-CM/PCS codes, dates of service, and payment information. Each record on this file contains the information from one health care claim. Episodes of care may encompass more than one health care claim.

2.5.2 Encounter DME File

The Encounter DME File contains health care encounters reported to CMS by MAOs in a format similar to the FFS DME claims but encounter records do not include payment information. Additionally, chart review records are included on this file and are a special type of MA encounter data that allows MAOs to add or remove diagnoses initially reported on encounter data records. The Encounter DME File may include encounter records reported for additional DME services provided by MA plans not covered by FFS Medicare. Episodes of care may encompass more than one health care encounter.

2.6 Home Health Agency (HHA) Files

2.6.1 Fee-for-Service HHA File

The FFS HHA File contains Medicare Part A final action claims submitted by HHA providers for reimbursement of home health covered services. Information in this file includes the number of visits, type of visit (skilled nursing care, home health aides, physical therapy, speech therapy, occupational therapy, and medical social services), for ICD-10-CM/PCS codes, revenue center codes, dates of service, and payment information. An HHA claim may cover services provided over a period of time, rather than a single day. Each record on this file contains the information from one health care claim. Episodes of care may encompass more than one health care claim.

2.6.2 Encounter HHA File

The Encounter HHA File contains health care encounters reported to CMS by MAOs in a format similar to the FFS HHA claims but encounter records do not include payment information. Additionally, chart review records are included on this file and are a special type of MA encounter data that allows MAOs to add or remove diagnoses initially reported on encounter data records. An HHA Encounter record may cover services provided over a period of time, rather than a single day. The encounter HHA File may include encounter records reported for additional HHA services provided by MA plans not covered by FFS Medicare. Episodes of care may encompass more than one health care encounter.

2.7 Hospice File

The Hospice File contains Medicare Part A final action claims data submitted by hospice providers. The data in this file include the type of hospice care received (e.g., routine home care or IP respite care). The Hospice File contains data fields for ICD-10 diagnosis codes, revenue center codes, dates of service, payment information, and some demographic information (such as date of birth, race, and sex). All Medicare beneficiaries receiving hospice care receive this benefit through Medicare FFS coverage, regardless of their type of Medicare enrollment (FFS or MA). Therefore, there is no separate Encounter Hospice file. Each record on this file contains the information from one health care claim. Episodes of care may encompass more than one health care claim.

3 Medicare Provider Analysis and Review (MedPAR) File

The MedPAR File contains IP hospitalization and SNF stays that were covered by FFS Medicare. MedPAR records are created by rolling up individual IP and SNF FFS claims for a single IP or SNF stay record. Each MedPAR record includes ICD-10 diagnosis and procedure codes associated with each IP or SNF stay. All Medicare Part A short-and long-stay hospitalization claims and SNF claims for each calendar year are included in the MedPAR file. Inclusion of hospital stay records on the MedPAR file are based on year of discharge. SNF stays are included based on year of admission into the facility.

4 Medicare Part D Prescription Drug Event (PDE) File

The Part D PDE File contains a summary of prescription drug claims submitted by pharmacies to Part D plan providers and payment data used by CMS to administer benefits for Medicare Part D enrollees, including payments to the Part D plan providers. Each record on this file includes the National Drug Code (NDC), days' supply, dates of service, and drug cost and payment information. It does not contain individual prescription drug claims, but rather summary records submitted to CMS by Medicare Part D prescription drug plan providers. The Medicare Part D PDE file contains one record for each prescription drug event. This file can contain multiple records per person.

Appendix II Detailed Description of Linkage Methodology

1 NHCS and CMS Medicare Linkage Submission Files

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

- SSN (validated)[8]
- DOB (month, day, and year)
- Sex
- Zip Code and State of residence
- First, middle initial, and last name

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

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

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

- Sex was missing. Two alternate records (one with male sex and the other with female) were created.
- Improbable date of birth. Medicare records with year of birth prior to 1903 were deleted.
- State of residence outside of U.S. and not in rest of world (RW) list. Alternate record was created with state code changed to missing
- Multiple name parts and common nicknames (see below)

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

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

Table I. Example of Alternate Record Generation using Name Fields

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

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

Submission files, which combined the cleaned and validated PII fields, were created separately for NHCS patient records and for CMS Medicare administrative records. During this process, multiple submission records were created for each patient/Medicare beneficiary to show all combinations of the recorded values for these fields. That is, if a patient/beneficiary 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/beneficiary (see Table II for example). Submission records that did not meet the eligibility requirements (see Section 3.1 Linkage Eligibility Determination) were removed from the submission file.

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

NHCS 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 next step in the linkage process, used only the NHCS and CMS

Medicare submission records that included a valid SSN. After records had been linked using SSN, the algorithm validated the deterministic links by comparing first name, middle initial, last name, month of birth, day of birth, year of birth, ZIP code of residence, and state of residence. If the ratio of agreeing identifiers to non-missing identifiers was greater than 50%, the linked pair was retained as a deterministic match. The collection of records resulting from the deterministic match is referred to as the 'truth source.'

3 Probabilistic Linkage

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

3.1 Blocking

Blocking is a key step in the probabilistic record linkage process. It identifies a smaller set of potential candidate pairs, eliminating the need to compare every single pair in the full comparison space (i.e., the Cartesian product). According to Christen, blocking or indexing, "splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key)."[9] Intuitively developed rules can be used to define the blocking criteria, however, for this linkage, variable values in the data being linked were used to inform the development of a set of blocking passes that efficiently join the datasets together (i.e., multiple, overlapping blocking passes are run, each using a different blocking key). By using these data to create an efficient blocking scheme (or set of blocking passes), a high percentage of true positive links were retained while the number of false positive links was significantly reduced. A supervised machine learning algorithm used the 'truth source' (see Appendix II section 2) as the validation dataset and the NHCS and CMS Medicare submission records as training data. For more detailed information on the supervised machine learning algorithm used, please refer to "Learning Blocking Schemes for Record Linkage" and "Using supervised machine learning to identify efficient blocking schemes for record linkage".[10] [11]

The machine learning algorithm produced 6 blocking passes to be used in the blocking scheme. Table III 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.

[9] 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 December 2025).
[10] 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 December 2025).
[11] 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 2025).

Table III. Blocking and Scoring Scheme Used to Identify and Score Potential Links

Key Number	Blocking Key	Scoring Key
1	Sex, day of birth, month of birth, year of birth, zip code of residence	First name, middle initial, last name*
2	First name, last name, year of birth	Middle initial, sex, day of birth, month of birth, zip code of residence, state of residence
3	First name, sex, day of birth, month of birth, year of birth	Middle initial, last name, zip code of residence, state of residence
4	Last name, sex, day of birth, month of birth	First name, middle initial, year of birth, zip code of residence, state of residence
5	Sex, day of birth, month of birth, year of birth, state of residence	First name, middle initial, last name, zip code of residence
6	Sex, month of birth, year of birth, zip code of residence, state of residence	First name, middle initial, last name, day of birth

* State was not used as a scoring variable for Block 1 since pairs were blocked on zip code, and state would not contain additional information on match validity.

3.2 Score Pairs

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

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

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

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

Except for first and last name, agreement status was set to 1 if the NHCS and Medicare values for a particular PII variable agreed exactly, 0 if they disagreed, and missing (i.e., '.') if either value was missing on the paired records. The agreement status assignment for first and last name is explained further in section 3.2.2 of this appendix.

3.2.1 M and U Probabilities

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

Table IV. Example of Agreement Flags Using Blocking Pass 1

Person identifier: NHCS Patient ID	Person identifier: Medicare Beneficiary ID	PII Agreement flag[1]: Middle Initial	PII Agreement flag[1]: Last Name	PII Agreement flag[1]: First Name
1	1	1	0	.
1	1	.	1	0
1	1	1	0	0
2	2	1	0	0
3	789	1	1	1
3	789	0	1	1
3	789	.	1	.
3	789	0	0	1
3	322	1	0	1

NOTES: Data have been fabricated for the purposes of this example. PII – Personally Identifiable Information.
[1] Agreement status of 1 = match, 0 = non-match, and . = missing values

Table V. Example Showing Summarization of Blocked Record Pairs for M-Probability Estimation, based on Table IV Example

Person identifier: NHCS Patient ID	Person identifier: Medicare Beneficiary ID	PII Agreement flag[1]: Middle Initial	PII Agreement flag[1]: Last Name	PII Agreement flag[1]: First Name
1	1	1	1	0
2	2	1	0	0
3	789	1	1	1
3	322	1	0	1

NOTES: Data have been fabricated for the purposes of this example. PII – Personally Identifiable Information.
[1] Agreement status of 1 = match, 0 = non-match, and . = missing values

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

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

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

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

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

For example, if for blocking pass 2, the state of residence code for FL appeared in 30,000 record pairings, agreed on 1,560 of those pairs, and the agreement rate for state of residence exceeded the 5th percentile, then the U-probability for Florida would have been computed as 1,560/30,000=0.052 or 5.2%. A 'catch-all' category was created for all identifier values that did not meet the above criteria. The U-probability of the 'catch-all' category was computed by dividing the total number of record pairs that agreed by the total number of record pairs being used to estimate the 'catch-all' category. The process for calculating U-probabilities for first and last name differs from these methods and is described in Section 3.2.2.

3.2.2 M and U Probabilities for First and Last Names

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

Because of the large number of unique name values, it was impractical to compute name specific U-probabilities for each blocking pass (i.e., there would not be enough records available for it to be done accurately). Instead, U-probabilities were estimated using pairs generated by the Cartesian product of all records in the NHCS linkage submission file and a simple random sample of 3% of records with non-missing name information from the Medicare submission file (approximately 3.0 million records for first name and 3.1 million records for last name).

Complete name tallies (separately, for first and last names) were then produced for the NHCS submission files. For each level of "common" name (defined as names appearing more than 100 times) on the survey submission file, 100,000 names were randomly selected from the Medicare submission file 3% sample for comparison. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. For each Jaro-Winkler level, the number of names in agreement of the 100,000 randomly selected Medicare file names were then tallied. A 'catch-all' category was calculated for the remaining "rare" names, based on 5 million name pairs generated by randomly sampling from both the list of rare names on the survey file and from the Medicare submission file 3% sample.[12] [13] [14]

[12] 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.
[13] 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.
[14] Resnick, D., Mirel, L., Roemer, M., & Campbell, S. (2020). Adjusting Record Linkage Match Weights to Partial Levels of String Agreement. *Everyone Counts: Data for the Public Good*. Joint Statistical Meetings (JSM). https://ww2.amstat.org/meetings/jsm/2020/onlineprogram/AbstractDetails.cfm?abstractid=312203 (accessed December 2025).

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 identifiers that had agreeing values. Similarly, non-agreement weights were only assigned to identifiers that had non-agreeing values. A non-agreement weight was always a negative value and reduced the pair weight score. It is important to note that if the M-probability was smaller than the U-probability (i.e., M<U), the pair score (see Section 3.2.4) was not adjusted according to the agreement/non-agreement weight. Because of the logarithmic function, having an M-probability that is smaller than the U-probability would have an inverse effect on the identifier agreement weights. That is, an agreement weight computed using a M-probability that was smaller than the U-probability would produce a negative weight, while the non-agreement weight would be positive. For example, if the M-probability for month of birth was 0.989 and the U-probability was 0.9999 then the agreement and non-agreement weights would be as follows,

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

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

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 followed the same general process:

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

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

continued two more times: for the 0.95 and 1.00 thresholds.

3.3 Probability Modeling

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

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

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

1. A pair-weight adjustment was computed (Adj_B) specific to blocking pass, B, by taking the log base 2 of the estimated number of matches (within blocking pass B) divided by the estimated number of non-matches in the blocking pass. For convenience, the estimated number of matches, $\widehat{N_{matches,B}}$, 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.[15]

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

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

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

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

When the last four digits of SSN did not agree:

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

No adjustment was made for pairs that did not have an SSN on either the NHCS patient or the CMS Medicare submission record. So, for these pairs:

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

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

4.1 Estimating Linkage Error to Determine Probability Cut-off Value

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

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

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

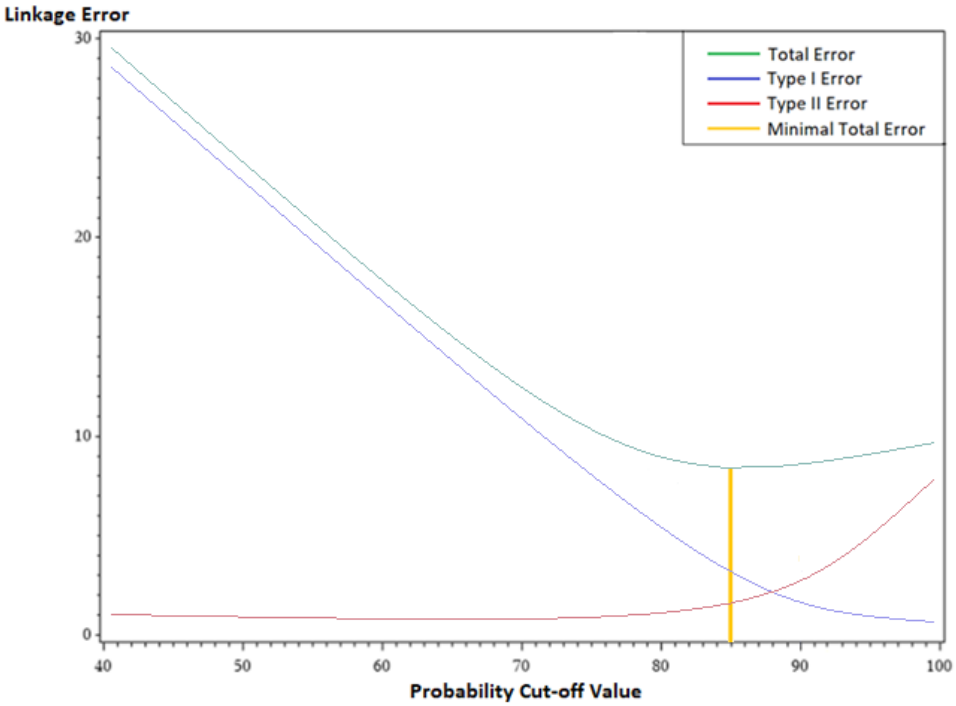
To measure Type II error, the truth source comprised of all matches identified in the deterministic linkage was used. Recall, the truth source contains records with full nine-digit SSN agreement (step 1) or with the last four digits of SSN in agreement (step 2). Potential deterministic matches were then validated using the available PII (see, Appendix II section 2). It was expected that this truth source had only a few exceptional pairs that were not true matches. For the probabilistic records, Type II error was estimated as the percentage of the truth source records that were not returned as links by the probabilistic method. Similarly to the computation of Type I error, an adjustment was made to the Type II error since some links having agreeing SSNs were being linked deterministically even if they were not returned by the probabilistic approach. For example, say that the probabilistic approach was able to

return 97% of true matches as links. If only a probabilistic linkage was conducted, the Type II error would then be 3%. However, among the 3% not linked probabilistically, some pairs could be linked deterministically. If the deterministic linkage rate is 50% (and if we assume the same rate among the non-linked pairs), then the Type II error rate can be estimated as 0.5*(1 – 0.97) = 0.015 or 1.5%.

4.2 Set Probability Cut-off Value

One goal of record linkage is to have the lowest errors possible. However, as more pairs are accepted, pairs that are less certain to be matches but accepted as links increase the Type I error and decrease Type II error. And as less pairs are accepted, pairs that are more certain to be matches but not accepted as links decrease the Type I error and increase Type II error. The optimal trade-off between Type I error and Type II error is not known, but it can be assumed to be optimal when the sum of Type I and Type II error is at a minimum. For this reason, Type I and Type II error are estimated at various probability cut-off values and the one that showed the lowest estimate of total error is selected (see Figure 1 for a stylized example). For the linkage of the NHCS and CMS Medicare data, the optimal probability cut-off value was set to 0.85.

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



4.3 Select Links Using Probability Cut-off Value

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

All record pairs with an adjusted probability value that fell below the cut-off (i.e., 0.85) were not linked.

4.4 Computed Error Rates of Selected Links

Final error rates were computed for selected links (described in Section 4.3). Table VI 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 2019 NHCS – 2019-2020 Medicare and 2020 NHCS – 2020-2021 Medicare 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 Medicare administrative record. In other words, if a link had an estimated probability of 0.98, then it was understood that there was a 98% chance this was a match. To estimate the Type I error rate for the probabilistic links, the chance that a link is not a match was summed (i.e., $\sum 1 - Probvalid_{SSN_{Adj}}$) and then divided by the total number of probabilistic records. The method to measure the overall Type II error remained unchanged (see Section 4.1).

Table VI. Algorithm Results for Total Selected Links

	Probability Cut-off Value	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
2019 NHCS	0.85	716,478	225,812	490,666	0.05%	0.43%
2020 NHCS	0.85	872,824	228,115	644,709	0.05%	0.49%