The Linkage of the 2016 National Hospital Care Survey to the 2016/2017 Centers for Medicare & Medicaid Services Medicare Enrollment, Claims/Encounters and Assessment Data: Matching Methodology and Analytic Considerations

Data Release Date: April 2024

Document Version Date: April 23, 2024

Division of Analysis and Epidemiology

National Center for Health Statistics

Centers for Disease Control and Prevention

datalinkage@cdc.gov

Suggested Citation: National Center for Health Statistics. Division of Analysis and Epidemiology. *The Linkage of the 2016 National Hospital Care Survey to the 2016/2017 Centers for Medicare & Medicaid Services Medicare Enrollment, Claims/Encounters and Assessment Data: Matching Methodology and Analytic Considerations*, April 2024. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data-linkage/index.htm

Contents

List of Acronyms

AMA, American Medical Association

BIC, Beneficiary Identification Code

CCD, Consolidated Clinic Document

CMS, Center for Medicare & Medicaid Services

DME, durable medical equipment

DOB, date of birth

ED, emergency department

EDB, enrollment database

EDRD, end-stage renal disease

EHR, electronic health record

ERB, ethics review board

FFS, fee for service

HCPCC, health care procedure classification code

HHA, home health agency

HICN, Health Insurance Claim Number

HMO, health maintenance organization

ICD-10-CM/PCS, International Classification of Diseases, 10th edition, Clinical Modification/Procedure Classification System

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

MDS, Minimum Data Set

MedPAR, Medicare Provider Analysis and Review File

NCHS, National Center for Health Statistics

NDI, National Death Index

NHCS, National Hospital Care Surveys

OASIS, Outcome and Assessment Information Set

OP, outpatient

OPD, outpatient department

PDE, prescription drug event

PDP, prescription drug plan

PII, personally identifiable information

PPO, preferred provider organization

RDC, Research Data Center

ResDAC, Research Data Assistance Center

SAF, standard analytic file

SNF, skilled nursing facility

SNP, special needs plan

SSN, social security number

UB, uniform billing

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. In addition to collecting and disseminating the Nation's official vital statistics, NCHS conducts several population-based surveys and healthcare establishment surveys, including the National Hospital Care Survey (NHCS), https://www.cdc.gov/nchs/nhcs/index.htm (accessed August 18, 2020). The 2016 NHCS sampled 581 hospitals, of which 158 agreed to participate and provided patient-level encounter records. Participating hospitals are requested to send all patient ambulatory care and inpatient (IP) encounters or claims records occurring within the data collection calendar year. The NHCS includes detailed information about each participating hospital's patients' characteristics, conditions, and treatment. Even though NHCS is an establishment survey (i.e., hospitals are the sampling unit) it collects patient personally identifiable information (PII), which enable data linkages.

Through its Data Linkage Program, NCHS has been able to expand the analytic utility of the data collected from NHCS by augmenting it with Medicare data collected from the Centers for Medicare & Medicaid Services (CMS). This report will describe the linkage of the 2016 NHCS to 2016/2017 CMS Medicare Data. Although NHCS is not currently nationally representative due to low response rates, 158/581=27%, linking NHCS with the CMS Medicare Data creates new data resources that can support a wide variety of health care services research projects.

This report includes a brief overview of the data sources, a description of the methods used for linkage, and analytic guidance to assist researchers when using the files. Detailed information on the linkage methodology is provided in Appendix I: Detailed Description of Linkage Methodology, and detailed descriptions of the Medicare files are described in Appendix II.

The data linkage work was performed at NCHS under contract #HHS2002016F92236B by NORC at the University of Chicago with funding from the Department of Health and Human Services' Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF).

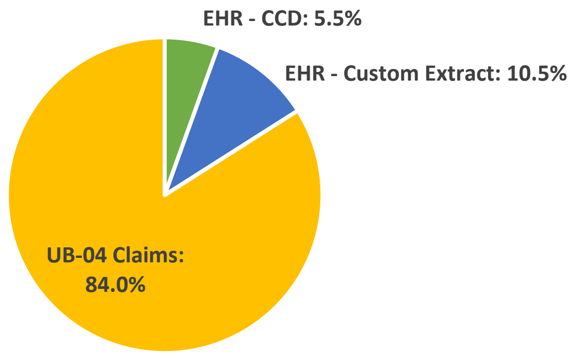
2 Background on Linked Files

2.1 National Hospital Care Survey

The NHCS is an establishment survey that collects IP, emergency department (ED), and outpatient department (OPD) episode-level data from sampled hospitals. NHCS is one of the NCHS National Healthcare Surveys, a family of surveys covering a wide spectrum of healthcare delivery settings from ambulatory and OPD to hospital and long-term care providers. 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 ambulatory care visits occurring during the calendar year. In previous years of the survey, hospitals were required to provide data from claims records, but to reduce the burden of reporting on participating hospitals, for the 2016 data collection hospitals were given the option of providing their data in the form of electronic health records (EHRs) or as claims records. Thus, participating hospitals provided data in the form of Uniform Bill (UB)-04 administrative claim records or EHR data, where the EHR data are provided in the form of Consolidated Clinic Documents (CCDs) or custom extracts. A distribution of the types of records received for the 2016 NHCS is provided in Figure 1. NHCS collects patient PII (e.g., full name, date of birth, and Social Security Number (SSN)), which allows for the linkage of episodes of care across hospital units as well as to other data sources, such as the CMS Medicare Data Files. The linkage described throughout this document only includes the linkage to CMS Medicare data for patients with either IP or ED visits – patients that only had other, non-ED OPD visits have been excluded.

Figure 1. Distribution of types of records received for the 2016 NHCS (only IP or ED visits) for total number of encounters



2.2 Centers for Medicare & Medicaid Services, Medicare Data

The 2016 NHCS has been linked to CMS Medicare enrollment, claims/encounters, and assessment data from 2016–2017.

Medicare is the primary 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). In 2016–2017, nearly two-thirds of persons enrolled in Medicare, known as Medicare beneficiaries, were enrolled in traditional Medicare, also known as Medicare fee-for-service (FFS). Nearly all Medicare FFS beneficiaries receive Part A hospital insurance benefits, which help cover IP hospital care, Skilled Nursing Facility (SNF) stays (not custodial or long-term care), home health care, and hospice care. Most FFS beneficiaries also enroll in Medicare Part B medical insurance benefits, which help to cover physician services, OP care, durable medical equipment (DME), and some home health care services.

In 2016–2017, approximately one-third of Medicare beneficiaries received Medicare benefits through a Medicare Advantage (MA) plan, also known as Medicare Part C. MA plans are administered by approved Medicare Advantage Organizations (MAOs). MAOs sponsor privately managed care plans such as Health Maintenance Organization (HMOs), Preferred Provider Organization (PPOs), and Special Needs Plans (SNPs) which provide, at a minimum, the same covered services provided in Medicare Parts A and B. MAOs may also elect to provide additional services not covered by Medicare Parts A and B such as dental and vision care. MAOs are responsible for providing Medicare benefits directly to enrollees through prior arrangements with providers or by paying for the benefits on behalf of enrollees.

In 2006, Medicare beneficiaries could begin to 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). MA-PDs provide prescription drug coverage that is integrated with the health care coverage provided to Medicare beneficiaries enrolled in MA plans.

The CMS Medicare Data Files are comprised of Standard Analytic Files, or SAFs, containing standard format extracts of research-oriented Medicare program data. The CMS Medicare Data Files 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 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, institutional outpatient (OP), practitioner/provider services (Carrier), home health agency (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, 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.

In addition to the SAFs and the PDE Files, two assessments are also included in the linked dataset – the Home Health Outcome and Assessment Information Set (OASIS) and the Long-Term Care Minimum Data Set (MDS). The OASIS assessment contains data pertaining to patient outcomes and home health care. The OASIS assessments are required of all HHAs certified to accept Medicare and Medicaid payments. The MDS is a health status screening and assessment tool used for all residents of long-term care nursing facilities certified to participate in Medicare or Medicaid, regardless of payer. The MDS assessment is also required for Medicare payment of SNF stays.

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

3 Linkage Methodology

3.1 Linkage Eligibility Determination

The linkage of 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' Research Ethics Review Board (ERB) and the linkage was performed only for linkage-eligible NHCS participants.[1]

Linkage was attempted only for NHCS patient records that met certain criteria (i.e., minimum levels of PII being available). The minimum criteria for a record to be considered linkage-eligible was that it had at least two of the following three identifiers present: valid SSN[2], valid date of birth (month, day, and year)[3] or complete name (first name, middle initial, and last name)[4]. For example, if the PII on the NHCS patient record had no SSN, a full name, and only the year of birth, the record would be considered ineligible for linkage, as only one of the criteria (i.e., that for name) was met.

The linkage eligibility status (which indicates whether the linkage eligibility criteria had been met) for a record is shown by the variable **ELIGSTAT.** The available values include 0 (ineligible) or 1 (eligible). Of note, only eligible patient records that match to a CMS enrollment record are included on the linked NHCS – CMS Medicare Data file. A supplementary file containing all NHCS patient IDs and **ELIGSTAT** was created. Users will be able to ascertain the total number of NHCS patients that were not eligible for linkage by using this file. 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 steps that were used to link the 2016 NHCS data with 2016/2017 CMS Medicare Enrollment Database (EDB). For more detailed information on linkage methodology (see Appendix I).

NHCS patient records were linked to records in the CMS EDB using the following identifiers: SSN, Medicare Health Insurance Claim Number (HICN), 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.

The 2016 NHCS patient records and the CMS EDB records were linked using both deterministic and probabilistic approaches. For the probabilistic approach, weighting was conducted according to the Fellegi-Sunter method.[5] Following this, a selection process was implemented

[1] The NCHS ERB, also known as an Institutional Review Board or IRB, is an appointed ethics review committee that is established to protect the rights and welfare of human research subjects.

[2] 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 cannot be 012345678

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

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

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

with the goal of selecting pairs believed to match (i.e., representing the same individual between the data sources). Table 1 highlights the linkage results by age, based on the following three steps (explained in further detail in Appendix I):

1. Deterministic linkage joins records on exact SSN or HICN, with links validated by comparing other identifying fields
2. Probabilistic linkage identified likely matches, or links, between all records. All deterministic matched pairs (from Step 1) were assigned a probabilistic match probability of 1; other records were 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 which were believed to represent the same individual between data sources

For each NHCS record that was linked, CMS extracted data records from its SAFs and provided them to NCHS.

Table 1. Linked 2016 NHCS – 2016/2017 CMS Medicare Administrative Records - Sample Sizes and Percent Linked, by Age

Age[1]	Sample Size: Total Sample	Sample Size: Eligible for Linkage[2]	Sample Size: Linked to 2016-2017 Medicare Administrative Data[3]	Percent Linked: Total Sample[4]	Percent Linked: Eligible Sample[5]
<65	3,692,926	3,459,122	294,388	8.0	8.5
>=65	762,766	717,624	699,734	91.7	97.5
Total	4,455,692	4,176,746	994,122	22.3	23.8

NOTES: Data are presented at patient level. It is possible that NHCS patients had more than one date of birth. When more than one date of birth was present, the minimum of the non-missing DOB was selected for the patient.

[1] Age is as of final encounter (date of last known contact). Age could not be determined for 24,121 patients based on availability of date of birth and age could not be determined for an additional 1,343,352 patients due to patient records missing PII.

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

[3] This group includes linkage-eligible patients who linked to Medicare administrative records at any time during the linkage interval (2016 - 2017).

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

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

4 Analytic Considerations

This section summarizes some key analytic issues for users of the linked NHCS data and CMS Medicare records. It is not an exhaustive list of the analytic issues that researchers may encounter while using the linked 2016 NHCS-2016/2017 CMS Medicare Data. This document will be updated as additional analytic issues are identified and brought to the attention of the NCHS Data Linkage Team (datalinkage@cdc.gov). Users of the linked NHCS - Medicare data Files are encouraged to visit the ResDAC website http://www.resdac.org (accessed August 18, 2020) for additional information on Medicare data and their analytic considerations.

4.1 Analytic Considerations for Linked NHCS Data

4.1.1 NHCS Hospital Eligibility and Sampling

Eligible hospitals for NHCS are non-institutional, non-federal hospitals with six or more staffed IP beds, and there are 6,622 hospitals which met these criteria as of 2013 to form the survey frame. A base sample of 500 hospitals and a reserve sample of 500 additional hospitals were drawn from this frame, within strata defined by bed size, urbanicity, and hospital type. Initially, the base sample of 500 hospitals was fielded. In 2013, to provide estimates for substance-involved ED visits, 81 hospitals with 500 staffed IP beds or more were added from the reserve sample. Thus, the hospital sample size for the 2016 NHCS data collection (which re-uses the 2013 sample) was 581 hospitals. In 2016, 158 out of the 581 sampled hospitals provided data and of the 158, 142 hospitals were eligible for linkage (note: this number excludes hospitals that did not provide patient PII and/or did not provide patient records covering at least 6 months of the analysis period). Of those 142 participating hospitals, 140 hospitals sent IP data and 121 hospitals sent ED data (i.e. 119 hospitals sent both IP and ED data, 21 hospitals sent IP data only, and 2 hospitals sent ED data only).

4.1.2 NHCS Sampling Weights Are Currently Not Available

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

4.1.3 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, ED, or OPD 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.[6]

4.2 Analytic Considerations for Linked Medicare Data Files

The 2016 NHCS patient-level records have been linked to the following CMS Medicare Data Files, which include enrollment data from the MBSF, claims/encounter data from the FFS and MA files, and patient assessment data from nursing home and home health care providers. The MBSF includes three segment files: the Base (Medicare Parts A/B/C/D), Cost & Utilization, and Chronic Conditions. More detailed descriptions of the linked Medicare data files listed in Table 2 are provided in Appendix II. The following sections address potential analytic considerations specific to each of the linked Medicare data files.

Table 2. List of 2016/2017 CMS Data Files linked to the 2016 NHCS survey file

CMS Data Files	Years
Master Beneficiary Summary File (MBSF)	2016-2017
Medicare Fee-for Service (Claim Files)	
Inpatient (IP)	2016-2017
Skilled Nursing Facility (SNF)	2016-2017
Professional (Carrier)	2016-2017
Outpatient (OP)	2016-2017
Durable Medical Equipment (DME)	2016-2017
Home Health Agency (HHA)	2016-2017
Hospice	2016-2017
Medicare Provider Analysis and Review File (MedPAR)	2016-2017
Medicare Advantage (Encounter Files)*	
Inpatient (IP)	2016-2017
Skilled Nursing Facility (SNF)	2016-2017
Professional (Carrier)	2016-2017
Outpatient (OP)	2016-2017
Durable Medical Equipment (DME)	2016-2017
Home Health Agency (HHA)	2016-2017
Medicare Part D Prescription Drug Event (PDE)	2016-2017
Home Health Outcome and Assessment Information Set (OASIS)	2016-2017
Long Term Care Minimum Data Set (MDS)	2016-2017

* At the time of the 2016 NHCS linkage to CMS Medicare Data, the 2017 Medicare Encounter Files were not available. The initial data release in September 2020 contained Medicare Advantage (MA) files for 2016 only. Medicare Advantage files for 2017 were added in the April 2024 data release.

[6] For more information on Patient_ID generation, see Technical Notes on page 14: https://www.cdc.gov/nchs/data/nhsr/nhsr097.pdf (accessed August 18, 2020)

4.2.1 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 2016 NHCS and 2016 MBSF will contain information for Medicare enrollment and summary health care utilization occurring in 2016.

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.

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

Creating Medicare Study Denominators

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 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 X, N, 0, or *, 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. It is possible to be enrolled in Medicare but not utilize Medicare services during the coverage period for a given calendar year.

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. This variable is coded by CMS using information provided by the Social Security Administration and/or Railroad Retirement Board. Knowing a beneficiary’s original reason for entitlement can be useful for identifying which aged beneficiaries were formerly entitled (i.e., prior to age 65) to Medicare due to a qualifying disability, since their cost and utilization profiles tend to differ from other aged beneficiaries, especially at ages 65-74. ENTLMT_RSN_ORIG values include: Old Age and Survivors Insurance (OASI), Disability Insurance Benefits (DIB) and 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 and RTI_RACE_CD. BENE_RACE_CD is the variable reported in the CMS administrative claims data system. The variable RTI_RACE_CD contains race and ethnicity codes imputed through the use of an algorithm developed by the Research Triangle Institute (RTI) and used by CMS to improve the accuracy of race and ethnicity data reported in the administrative claims data system. More detailed information regarding the RTI algorithm can be found at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4195038 (accessed August 18, 2020). Although patient race is reported in the NHCS data, the percent of patients with a survey reported valid race code is low. Researchers may wish to consider utilizing the race and ethnicity data present in the linked CMS administrative records.

4.2.1.2 MBSF Cost and Utilization Segment

The linked MBSF Cost and Utilization 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. The MBSF variables associated with FFS costs and payments may contain extreme outliers. Users may wish to consider applying top or bottom coding limits for these variables as these extreme values may adversely affect statistical calculations. Additional information about the variables included in the linked NHCS MBSF Cost and Utilization segment is available at https://www.resdac.org/cms-data/files/mbsf-cost-and-utilization (accessed August 18, 2020).

4.2.1.3 MBSF Chronic Conditions Segment

The CMS Medicare MBSF Chronic Conditions segment flags each Medicare FFS-enrolled beneficiary for the presence of one of 27 specific chronic conditions. Additional information about the methodology used to assign chronic condition flags to Medicare beneficiaries is available at https://www.ccwdata.org/web/guest/condition-categories (accessed August 18, 2020). 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, accessed August 18, 2020).

4.2.1.4 MBSF File Year Indicator

The MBSF reference year can be found in the variable BENE_ENROLLMT_REF_YR. Please note that both 2016 and 2017 linked records are appended into a single file. It is possible that a single beneficiary can have MBSF records for both 2016 and 2017. If this is the case, the beneficiary will appear twice in the file.

4.3 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 traditional 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 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 which can be accessed at www.resdac.org (accessed August 18, 2020) or www.ccwdata.org (accessed August 18, 2020).

A small number of FFS claims records may not have a corresponding MBSF record for that NHCS patient in that calendar year. There may be some record keeping inconsistencies because CMS data are collected for administrative, not research purposes. Data users may wish to exclude these records from their analytic sample.

4.3.1 Carrier File

The claims on the FFS Carrier File are processed by private carriers working under contract to CMS. Each carrier claim includes a Health Care Procedure Classification Code (HCPCS) to describe the nature of the billed service. The HCPCS are composed primarily of Level I HCPCS or 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. LINE_PLACE_OF_SRVC_CD is not a validated code and may contain inaccuracies.

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 4.3.2.

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. In the case of laboratory tests, the diagnosis will often be XX000, because the outside laboratory has no information from the physician about the reason for the test.

Some services may not appear in the Carrier claims, although they may have been received by the beneficiary. For example, CMS pays physicians a fixed amount for surgeries; this practice is called bundling. As part of bundling, CMS expects that certain care will be included in the payment amount, such as the first one or two office visits following surgery, or a biopsy just before surgery. Bundled services will not appear in the physician data. Interpretation of the rules on bundling varies by carrier.

4.3.2 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. Each year, CMS distributes a jurisdiction list, available from the CMS website, which specifies whether a carrier or a DMERC can process a claim for a particular service. Often, both carriers and DMERCs are allowed to process and pay a DME claims service, depending on whether the DME was provided as ‘‘incident to the physician’s service.’’

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.

4.3.3 Hospice File

All linked NHCS beneficiaries utilizing Hospice services in the Hospice File have a primary diagnosis, but most (90%) have no secondary diagnosis. Although data fields exist for procedure codes, such information generally is not reliable when recorded in hospice claims. 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 a FFS or MA plan.

4.3.4 Outpatient 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.

4.3.5 Inpatient 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 which summarizes individual inpatient claims at the stay level. (Section 4.3.7) 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 Outpatient File.

4.3.6 Skilled Nursing Facility (SNF) File

Each claim record on this file represents a health care claim submitted for payment by a skilled nursing facility 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 claim information summarized at the stay level may wish to use the MedPAR file which summarizes individual SNF claims at the stay level. (Section 4.3.7) 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.

4.3.7 Medicare Provider Analysis and Review (MedPAR) File

The MedPAR file was specifically developed by CMS to assist researchers interested in studying IP hospital and SNF care. 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. However, these claims will not be comprehensive, and 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 (accessed August 18, 2020). The CMS FFS IP and SNF Claims Files do not include “information only” claims.

The following fields on MedPAR Files are not used for payment purposes and should be used with caution:

- Source of admission (SRC_IP_ADMSN_CD)
  - This can include admissions due to transfers between facilities such as SNFs or other hospitals, admissions from the ED, and other referrals.
- Group health organization payment code (GHO_PD_CD)

In addition, MedPAR Files include a mortality variable. However, if the outcome of interest is mortality, users should request to use the mortality status from the 2016 NHCS Linked Mortality Files (accessed August 18, 2020).

At this time, CMS has not created a file similar to the MedPAR file for MA IP and SNF encounters; however, all individual IP and SNF encounter records submitted by the MAOs are available for analysis on the linked IP and SNF Encounter Data Files.

4.4 Analytic Considerations Specific to Medicare Advantage Encounter Files

MA encounter data reflect services provided to Medicare beneficiaries enrolled in MA plans, also known as Medicare Part C. There are important differences between MA encounter data and Medicare FFS claims data. 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. CMS primarily uses MA encounter data to help determine risk adjustment factors for each beneficiary, based on diagnoses reported in MA encounter records, which in turn are used to adjust CMS’ payments to MAOs. However, risk adjustment factors are only based on diagnosis data from IP, OP, and professional services (carrier) encounter records. CMS uses MA encounter data records for other purposes than risk adjustment including conducting quality review and improvement activities and other program oversight functions.

CMS acknowledges that while MA encounter data records most likely represent the majority but not all of health care services provided to MA enrollees, and due to differences in collection and payment purpose of MA encounter data, there may be differences in the completeness of encounter data versus FFS claims data. Generally, CMS MAOs are required to submit encounter data within 13 months after the end of the service calendar year. CMS has granted extensions of this deadline to help facilitate the submission of complete and accurate encounter data by MAOs.

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

Encounter data records record information on health care services provided to MA-enrolled beneficiaries. MA encounter records differ from FFS claims because they are: 1) 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 as MA plans may choose to offer additional services not covered by FFS Medicare, 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).

Record counts for 2016 NHCS Linked Medicare Encounter Files and the proportion of encounter records that are chart review records are provided in Table 3 below.

Table 3. 2016 NHCS linked to 2016 Medicare Encounter File record counts and proportions of records categorized as chart review

Encounter Type	Total Encounter File Record Count	Chart Review Record Count	% of Records that are Chart Review
Inpatient (IP)	279,742	87,619	23.9
Skilled Nursing Facility (SNF)	87,788	785	0.9
Home Health (HH)	279,567	1,191	0.4
Institutional Outpatient (OP)	2,069,756	30,937	1.5
Professional (Carrier)	13,629,073	1,178,283	8.0
Durable Medical Equipment (DME)	680,178	2,228	0.3

NOTES: Data are presented at record level. This table only represents 2016 encounter data as 2017 encounter data were unavailable at the time of linkage.

CMS has published additional guidance to assist with analysis of Medicare encounter claims data which can be accessed at http://www.resdac.org (accessed August 18, 2020) or https://www2.ccwdata.org/documents/10280/19002246/ccw-medicare-encounter-data-user-guide.pdf (accessed August 18, 2020).

A small number of MA encounter records may not have a corresponding MBSF record for that NHCS patient in that calendar year. There may be some record keeping inconsistencies because CMS data are collected for administrative, not research purposes. Data users may wish to exclude these records from their analytic sample.

4.5 Analytic Considerations Specific to the 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. All Medicare Part D prescription drug benefits are provided through private plans (plan sponsors).

Claims for prescription drugs are submitted by pharmacies to the Part D health plans for beneficiaries enrolled in Medicare Part D. PDE data are created by Part D health plans from point-of-service transactional data at the time a prescription is filled. Data for prescriptions that are ordered but not filled do not exist in this database. Not all Medicare-enrolled beneficiaries elect to purchase Part D coverage. Note that PDE data are not submitted by plans that receive retiree drug subsidies (RDS), or for other types of plans that are considered to be Part D creditable coverage (e.g., Veterans Administration [VA] or TRICARE).

PDE differs from a pharmacy claim in several ways. Each PDE record is a summary record containing the final status of a drug claim sent by a pharmacy to Part D sponsors, accounting for any subsequent adjustments. Pharmacy claims rejected by the sponsor are not included in PDE data. For example, if a pharmacy submits an original claim to a plan sponsor that is rejected due to a prior authorization requirement, and later, when the prior authorization criteria are met, resubmits the claim which is then accepted by the sponsor, the sponsor would then submit only one PDE record to CMS reflecting the final status of the accepted claim. Similarly, if a pharmacy submits a claim to a plan sponsor and then soon after reverses (cancels) the claim, the sponsor would not submit a PDE record to CMS. Additionally, since the PDE data represent ‘‘final action,’’ all PDE adjustments received by CMS through the PDE submission deadline for payment reconciliation is accounted for in the data, including PDE adjustments, resubmissions, and deletions.

Not all drugs used by Part D-enrolled beneficiaries are included in the PDE Files. PDE data generally do not include Part D-excluded prescription drugs (unless the MA-PD plan covers excluded drugs as a supplemental benefit). Prescriptions obtained through a third party (e.g., VA) or those for which a claim is not submitted (e.g., if a beneficiary pays cash out of pocket) are not available. In addition, over-the-counter (OTC) drugs are excluded from Part D and typically are not included in the PDE Files, unless they are part of an approved step therapy protocol.

CMS has published additional guidance to assist with analysis of Medicare prescription drug which can be accessed at http://www.resdac.org (accessed August 18, 2020) or https://www.ccwdata.org (accessed August 18, 2020).

5 Access to Data Files

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

To ensure confidentiality, NCHS provides safeguards including the removal of all personal identifiers from analytic linked files. Additionally, the linked data files are only made available in secure facilities for approved research projects. Researchers who want to access the linked 2016 NCHS- 2016/2017 CMS Medicare Data Files must submit a research proposal to the NCHS Research Data Center (RDC) to obtain permission to access the restricted use files. All researchers must submit a research proposal to determine if their project is feasible and to gain access to these restricted data files. The proposal provides a framework which allows RDC staff to identify potential disclosure risks. More information regarding RDC and instructions for submitting an RDC proposal are available from: https://www.cdc.gov/rdc/ (accessed August 18, 2020).

5.1 Merging NHCS Analytic Files to the NHCS-CMS Medicare Linked Data

NHCS is an establishment survey where the respondents are individual hospitals rather than their patients. Typically, this type of survey restricts analyses to the sample unit-level, but because NHCS collects hospital encounter-level records, encounter-level analysis is also possible. For NHCS patient with either an IP discharge or ED visit, results of the patient-level linkage to the CMS Medicare Data are available in the linked 2016 NHCS-2016/2017 CMS Medicare Data Files.

To perform NHCS patient encounter-level analysis, the linked 2016 NHCS-2016/2017 CMS Medicare Data Files can be used in conjunction with 2016 NHCS analytic files.[7] The variable PATIENT_ID allows analysts to merge NHCS patient records for the same patients within or across hospital settings (IP or ED) and to merge information from the NHCS-CMS Medicare Linked Data files. For additional information on how to link information across NCHS-CMS Medicare Linked Segment files see Appendix 2, Section 2.

Note: All RDC applications to analyze linked NHCS-CMS data should include requests to analyze the MBSF for the same calendar years as the Medicare health care claims, encounter, prescription drug data, or assessment data to allow researchers to determine the correct study denominators for the various Medicare programs. 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.

5.2 Additional Related Data Sources

5.2.1 Linked NHCS-CMS T-MSIS Files

Analysts interested in studying health care utilization and costs for the dually eligible population (persons enrolled in both Medicare and Medicaid) may wish to also request access to the 2016 NHCS–CMS T-MSIS linked data files (accessed 03/19/2024) for Medicaid enrollment and claims

[7] Find more information about the NHCS analytic files: https://www.cdc.gov/rdc/b1datatype/dt1224h.htm (accessed August 18, 2020)

data from 2015–2017. Medicare is the first payer for health care services covered by Medicare Parts A, B, C, and D, with Medicaid providing supplemental coverage for covered Medicare services including copayment and deductible amounts up to the limits identified in the state Medicaid plan. To integrate the linked NHCS-CMS T-MSIS linked data files into the linked NHCS–CMS Medicare data files, joins are made on the common identification number, PATIENT_ID.

More information about the linked 2016 NHCS-CMS T-MSIS data files can be found at: https://www.cdc.gov/nchs/data-linkage/nhcs-medicaid.htm (accessed March 19, 2024).

5.2.2 Linked NHCS-NDI Mortality Files

Analysts interested in studying mortality among the 2016 NHCS patient population enrolled in Medicare are encouraged to use the linked mortality data available in the 2016 NHCS- NDI Mortality files rather than the mortality data available in the linked 2016 NHCS-CMS T-MSIS files. The linked 2016 NHCS-NDI Mortality files (accessed March 22, 2024) include information on deaths identified for the entire 2016 NHCS patient population through linkage with the National Death Index and are not limited to deaths among the Medicare enrolled population. In addition, in the NHCS-NDI linked data cause of death is available for patients who died. The linked mortality file includes Patient ID, date of birth, date of death, and cause of death information for linked decedents. To integrate the linked NHCS-NDI linked data files into the linked NHCS-CMS T-MSIS data files, joins are made on the common identification number, PATIENT_ID.

More information about the linked 2016 NHCS-NDI Mortality data files can be found at: https://www.cdc.gov/nchs/data-linkage/nhcs-ndi.htm (accessed March 22, 2024)

5.2.3 Linked NHCS–Housing and Urban Development (HUD) Administrative Data Files

Researchers interested in outcomes related to housing insecurity may also request variables from the linked 2016 NHCS–2015-2017 HUD Administrative Data file if housing assistance is a variable/outcome of interest (Restricted-Use Linked NHCS – HUD Administrative Housing Data, accessed March 22, 2024). 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 into the linked 2016 NHCS-2016/2017 CMS Medicare data files, joins are made on the common identification number, PATIENT_ID.

5.2.4 Linked NHCS–Department of Veterans Affairs (VA) Data Files

Researchers interested in outcomes related to Veterans may also request variables from the Linked NHCS–VA administrative data files (accessed March 22, 2024). The Linked NHCS–VA data files include information on a wide range of health-related topics for Veterans, including Veteran status and utilization of VA benefit programs. To integrate the linked NHCS–VA linked data files into the linked NHCS–CMS Medicare data files, joins are made on the common identification number, PATIENT_ID.

Appendix I: Detailed Description of Linkage Methodology

1 NHCS and CMS Linkage Submission Files

Prior to the linkage of the NHCS and CMS administrative records, there were a series of processes that performed various data cleaning routines on the fields of these files: processing was conducted separately for NHCS and CMS records. Each of the listed PII fields was individually processed and output to its own table (i.e., there were separate tables for SSN, DOB, first name, etc., each record showing a possible value for that field for each patient or enrollee):

- SSN validation.[8]
- HICN[9]
- DOB
- Sex
- ZIP Code and State of residence
- First name, middle initial, and last name

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

- Date values: when invalid or outside of expected range, they are set to null
- Sex values: when multiple sex values are seen for the same person, sex is set to null
- Name values: multiple edits are applied:
    - Removal of special characters such as [“-.,<>/?, etc.]
    - Removal of descriptive words such as twin, brother, daughter, etc.
    - Nulling of baby names—it is common for hospitals to use the mother’s first name when no name has been decided for the baby
    - Nulling of Jane/John Doe
    - Removal of titles such as Mister, Miss, etc.
    - Removal of suffixes such as Junior, II, etc.
    - Removal of special text unique to survey such as first name listed as “Void”

Similar to the cleaning process, a more elaborate routine was used to generate alternate records involving the name fields. For patients with multiple name parts, additional records were generated using each individual piece as a possible name value. Table 4 below provides two examples of how name information was used to generate alternate records, using hypothetical data. For patient A, the first name was used to generate multiple records, and for patient B, the last name was used.

[8] SSN is considered valid if: 9-digits in length, containing only numbers, does not begin with 000, 666, or any values after 899, all 9-digits cannot be the same (i.e. 111111111, etc.), middle two and last 4-digits cannot be 0’s (i.e. xxx-00-xxxx or xxx-xx-0000), and is not 012345678

[9] HICN is considered valid if: 10-11 digits in length, first 9-digits contain only numbers and the last 1 or 2 digits contain a correct Beneficiary Identification Code (BIC)

Table 4. Example of Alternate Record Generation using Name Fields

Patient ID	First Name	Middle Initial	Last Name	Alternate Record
A	John H		Smith	0
A	John	H	Smith	1
A	H		Smith	1
A	John		Smith	1
B	John	R	Smith Jones	0
B	John	R	Smith	1
B	John	R	Jones	1

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

A submission file that combined the cleaned and validated patient PII was created for NHCS records and for CMS records. During this process, multiple submission file records were created for each patient/beneficiary to show all combinations of the recorded values for these fields. That is, if a patient had two states of residence recorded and three date-of-birth variants recorded and each of the remaining fields had only one variant, then six submission records would be created for this patient.

2 Deterministic Linkage Using Unique Identifiers

The deterministic linkage was the next step in the linkage process. The deterministic linkage used only the eligible NHCS and CMS records that were submitted with a valid format SSN or HICN. Linkage eligibility is defined earlier in this report (see Section 3.1 Linkage Eligibility Determination). In some cases, a valid SSN was extracted from a HICN. When the Beneficiary Identification Code (BIC) was identified as either A, J, M, or T, this indicated that the first 9 digits of the HICN were that beneficiaries' SSN. If a patient/beneficiary does not have a valid SSN or if the extracted SSN differs from an already cleaned SSN, the extracted SSN value is retained as an additional SSN value to be used in the linkage process.

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 matching 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 was referred to as the 'truth deck.'

3 Probabilistic Linkage

The second step in the linkage process was to perform the probabilistic linkage. 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 describes 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 it, each pair was assigned an estimated probability representing the likelihood that it is a match – using pair weights computed (according to formula) for each identifier in the pair – before selecting the most probable match between two records.

3.1 Blocking

Blocking is a key step in the probabilistic record linkage process. It identified a smaller set of potential candidate pairs without having 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)."[10] Intuitively developed rules can be used to define the blocking criteria, however, for this linkage, data 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 the data to create an efficient block scheme (or set of blocking passes), a high percentage of true positive links were retained while significantly reducing the number of false positive links. A supervised machine learning algorithm used the 'truth deck' as the validation dataset and a sample of the NHCS and CMS EDB records as the training dataset. For more detailed information on this method please refer to "Learning Blocking Schemes for /Record Linkage."[11]

The machine learning algorithm generated 6 blocking passes to be used in the blocking scheme. Tables 5 and 6 provides a specific breakdown of each blocking pass:

Table 5. Blocking variables used to identify linked records

Block Key 1	Block Key 2	Block Key 3	Block Key 4	Block Key 5	Block Key 6
• Day of birth • Month of birth • Year of birth • ZIP code	• First name • Last name • Month of birth • Year of birth	• First name • Day of birth • Month of birth • Year of birth • Sex	• Last name • Day of birth • Month of birth • Year of birth • Sex	• First name • Last name • State of residence	• Middle initial • Day of birth • Month of birth • Year of birth • State of residence • Sex

Table 6. Variables used to score linked records in each blocking pass

Score Blocking Pass 1	Score Blocking Pass 2	Score Blocking Pass 3	Score Blocking Pass 4	Score Blocking Pass 5	Score Blocking Pass 6
• First name • Middle initial • Last name • Sex	• Middle initial • Day of birth • State of residence	• Middle initial • Last name • State of residence	• First name • Middle initial • State of residence	• Middle initial • Day of birth	• First name • Last name • ZIP code

[10] Christen, Peter. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Berlin Heidelberg: Springer-Verlag, 2012. http://www.springer.com/us/book/9783642311635 (accessed August 18, 2020).

[11] Michelson, Matthew, and Craig A. Knoblock. "Learning Blocking Schemes for Record Linkage." In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, 440–445. AAAI'06. Boston, Massachusetts: AAAI Press, 2006. https://pdfs.semanticscholar.org/18ee/d721845dd876c769c1fd2d967c04f3a6eeaa.pdf (accessed August 18, 2020).

	• ZIP code • Sex	• ZIP code	• ZIP code	• Month of birth • Year of birth • ZIP code • Sex	

3.2 Score Pairs

Next, each pair was weighted 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 2.3), which allowed the linkage algorithm to select final links to include in the linked file. The scoring process followed the following order:

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

The pair scores were calculated on the 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 (conditional on sex) or Last Initial (when applicable)
- Year of Birth
- Month of Birth
- Day of Birth
- Sex
- State of Residence
- ZIP Code

3.2.1 Calculate M- and U- Probabilities

The **M-probability** – the probability that identifiers from the paired records agree, given that records represent the same person – were estimated separately within each individual blocking pass. M-probabilities were calculated for each of the identifiers not used in the blocking key (Table 6). Within the blocking pass, pairs with non-missing and agreeing (defined as 8 or more digits being the same) SSN were used to calculate the M-probabilities, as these are assumed to represent the same individual. For example, among qualifying pairs in blocking pass 2, 99.4% agree on day of birth and 94.5% agreed on state of residence. These percentages represented estimates of the M-probabilities for these identifiers.

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 the name field
- Jaro-Winkler Similarity Levels – this process is explained in greater detail in Section 3.2.2
- Last name is conditional on sex – because women frequently change their maiden name to their spouse's last name after marriage (or may change back to maiden in event of divorce/widowing), this resulted in a lower agreement last name M-probabilities for the female population, and was taken into consideration when computing corresponding agreement and non-agreement weights.

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

Similar to the M-probabilities, U-probabilities were only calculated for the non-blocking variables. However, for this linkage, the U-probabilities were calculated for each value (level) of a variable (i.e., the values/levels for state of residence are Pennsylvania, Florida, etc.). For example, the state of residence U-probabilities within blocking pass 2 for Florida and Pennsylvania were, 0.052 (5.2%) and 0.091 (9.1%), respectively. However, for first and last name, the U-probabilities were not calculated exactly in the same manner, and the method used for them is described in Section 3.2.2.

3.2.2 M and U Probabilities for First and Last Names

Similar to the M-probability, Jaro-Winkler levels (0.85, 0.90, 0.95, and 1.00) were calculated for use in the U-probability computation. Because agreement levels fall over a range, first and last name U-probabilities were computed for each Jaro-Winkler score level. Since there are a plethora of possible values for first and last name (i.e., one for each possible name), it was impractical to compute U- probabilities specific name for each blocking pass (i.e., there would not be enough records available for it to be done accurately). Instead, U-probabilities were estimated using pairs generated by the Cartesian product of all records in the NHCS submission file and a simple random sample of 1% (1,130,397 records for first name and 1,144,078 records for last name) of records with non-missing name information of the CMS Medicare EDB submission file.

Complete name tallies (separately, for first and last names) were then produced for the NHCS submission file. For each level of name on the file, 100,000 names were randomly selected from the CMS Medicare EDB submission file 1% sample to compare to it. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. The number of names in agreeance of the 100,000 randomly selected CMS Medicare EDB names that agreed at that level for each name were then tallied.[12,13]

[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.

3.2.3 Calculate Agreement and Non-Agreement Weights

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

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

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

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

3.2.4 Calculate Pair Weight Scores

In the next step, pair weights were calculated, which were then used in the probability model. The pair weights were calculated differently for each record pair, but follow the same general process:

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

First name and last name weights were assigned using Jaro-Winkler similarity scores described in Section 3.2.2. These scores ranged from 0 to 1, with 0 representing no similarity and 1 representing exact agreement. The weighting algorithm assigned all scores below 0.85 a disagreement weight. The algorithm assigned all scores above 0.85 an agreement weight associated with the 0.85 level. If there was an agreement at the 0.85 level, the algorithm assessed the pair at the 0.90 level *given* that it agreed at the 0.85 level. If the names disagreed at this level, the algorithm assigned them a disagreement weight (specific to the 0.90 level given agreement at the 0.85 level). If the names agreed, the algorithm assigned them an additional agreement weight (specific to the 0.90 level). This process continued two more times: for the 0.95 and 1.00 thresholds.

3.3 Probability Modeling

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

- Combine pairs across blocking passes (Pair-weights are specific to each blocking pass and are not comparable)
- Select a "best" record among patient's IDs that have linked to multiple Beneficiary IDs

- Select final matches based on a probability threshold (discussed in the following section)

The partial EM model was an iterative process that can be described in 4 steps:

1. A pair-weight adjustment (Adj_B) was computed 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 used in the first iteration was set to half of the pairs in the blocking pass (i.e., all pairs generated by the blocking pass specification). The number of non-matches was computed by subtracting the estimated number of matches from the number of pairs (regardless of how likely they are to be matches) in the blocking pass.

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

   Note that in the first iteration, it was assumed that the number of matches (within blocking pass *B*) were equal to the number of non-matches (within blocking pass *B*) resulting in $Adj_B = 0$. . If however, in a later iteration, the number of matches was estimated to be 20,000 and the number of pairs is 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*, were computed in blocking pass, *B*, being a match by taking 2 to the power of the adjusted pair-weight (sum of pair-weight (*PW*) and 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 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 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 the number of matches (within blocking pass 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.[14]

To remedy this, before the algorithm adjudicated the matches against the probability threshold, one final adjustment was made to the match probabilities (for probabilistic pairs). For pairs that had an SSN on both the NHCS and CMS EDB record, the estimated probability was adjusted based on the last four digits of the SSN.[15]

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

[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[16] agreed (i.e., are exactly the same):

$$Probvalid_{SSN_{Adj}} = \frac{\left(\frac{P_{EM}(Match)}{1 - P_{EM}(Match)} \cdot \frac{M_{SSN-L4}}{U_{SSN-L4}}\right)}{\left(\left(\frac{P_{EM}(Match)}{1 - P_{EM}(Match)} \cdot \frac{M_{SSN-L4}}{U_{SSN-L4}}\right) + 1\right)}$$

When the last four digits of SSN did not agree (and HICN did not agree):

$$Probvalid_{SSN_{Adj}} = \frac{\left(\frac{P_{EM}(Match)}{1 - P_{EM}(Match)} \cdot \frac{(1 - M_{SSN-L4})}{(1 - U_{SSN-L4})}\right)}{\left(\left(\frac{P_{EM}(Match)}{1 - P_{EM}(Match)} \cdot \frac{(1 - M_{SSN-L4})}{(1 - U_{SSN-L4})}\right) + 1\right)}$$

For pairs that did not have an SSN on either the NHCS or CMS EDB record, came from deterministic linkage, or which had last four digits of SSN disagreeing but HICN agreeing, no adjustment was made. So, for these pairs:

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

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

4.1 Estimating Linkage Error of Selected Links

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

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

The estimated probabilities were used to measure Type I error. For the probabilistic records, the estimated match probabilities represented the probability that the NHCS record was a match to the CMS EDB 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 correctly matched. To estimate the Type I error rate for the probabilistic links, the chance that a link is not a match was summed and then divided by the total number of probabilistic records. 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

[16] Rather than using the entire SSN, the last four digits are used since the first five digits of an SSN are not truly random. Prior to 06/25/2011 the first three digits represented the state where the SSA paperwork was submitted to obtain an SSN. The fourth and fifth digit are known as a group number that cycles from 01 to 99.[16] This additional pair weight allows for more accurate adjudication of links where other PII may not provide a clear indication of match status.

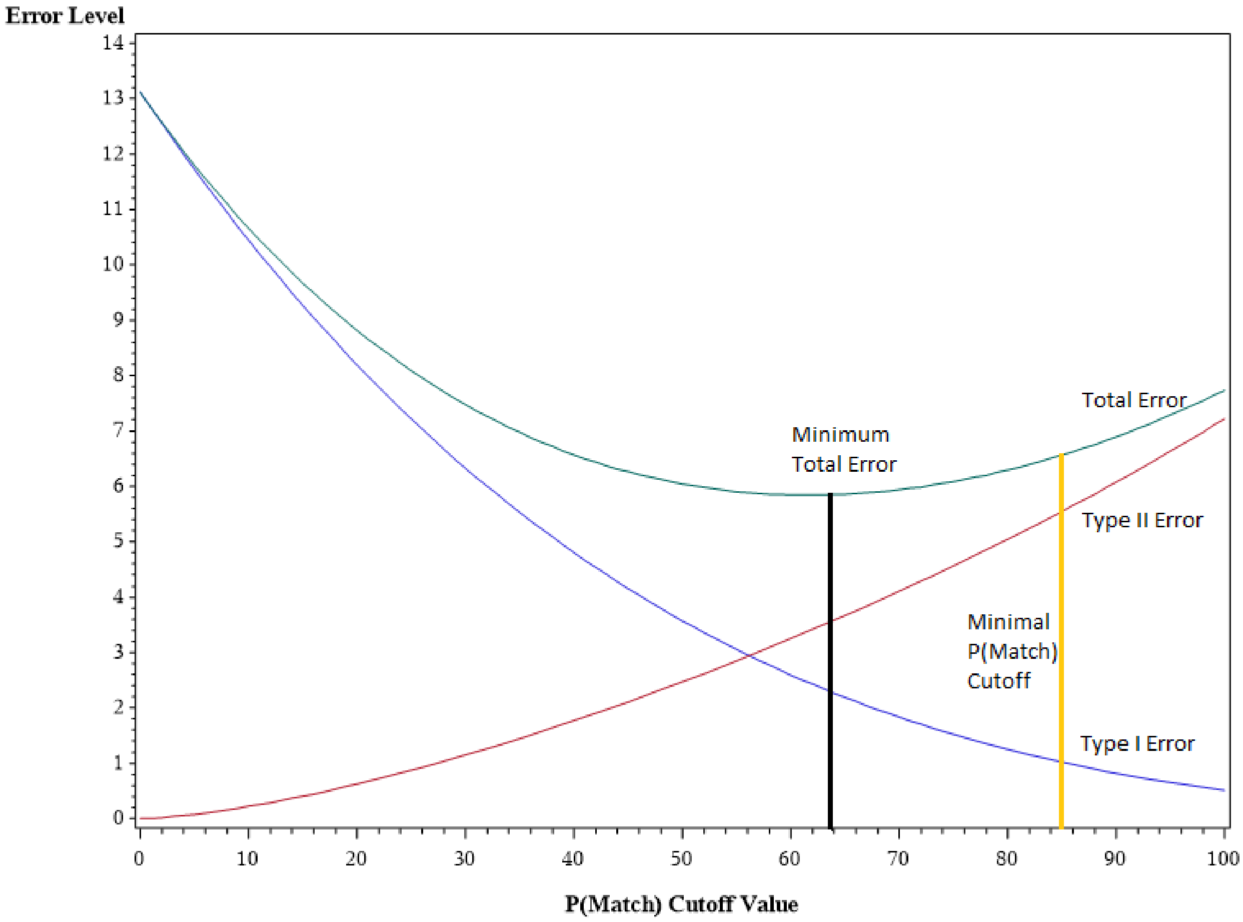
deterministically established links and so the estimate was likely biased low. Since a sizeable proportion of links were derived from the deterministic method, this had the effect of reducing the estimated Type I error by the proportion of probabilistically determined linkages among all linkages. For example, the Type I error rate was estimated for probabilistic links as 1.2%, but only 40% of all links were derived from probabilistic analysis. Thus, the estimated Type I error rate for the combined linkage process was (0.40*0.012) = 0.0048 or 0.48%.

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

4.2 Set Probability Cutoff

The goal of record linkage was to have the lowest errors possible. However, as more pairs were accepted, pairs that were less certain to be matches as links increase the Type I error and decrease Type II error (see Figure 2). And as less pairs were accepted, pairs that were more certain to be matches as links decrease the Type I error and increase Type II error. The optimal trade-off is between Type I error and Type II error was not known, and likely this depends on the type of analysis to be conducted with the linked data, but it is assumed that it is not far from optimality when the sum of Type I and Type II error is at a minimum. For this reason, Type I and Type II error are estimated at various probability cut points and the one that showed the lowest estimate of total error was selected. However, because there are concerns that using pairs with low P(Match) might be inappropriate for certain analyses of linked records, P(Match) = .85 was established as the lowest threshold that will be used for the acceptance of links into datasets made available for external researchers.

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



4.3 Select Links Using Probability Threshold

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

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

4.4 Computed Error Rates

Overall, the Type I and Type II linkage error rates for the 2016 NHCS – 2016/2017 CMS Medicare Data linkage were 0.02% and 0.21, respectively. Additionally, linkage error rates were assessed based on the type of record source (UB-04 claim, EHR custom extract or CCD). Table 7 provides 2016 NHCS patient linkage results for both UB-04 claims and EHRs. The table reports the finalized results with the probability cutoff threshold chosen by the algorithm. As noted in the table, EHRs have slightly higher estimated linkage error (both Type I and II) compared to the UB-04 claims records. Due to elevated levels of missing data in EHRs compared to the UB-04 claims records, the number of deterministic matches made by the algorithm for EHR Custom Extract (89.3%) is proportionally higher than UB-04 deterministic matches (77.2%). This resulted in a lower proportion of EHRs having CMS Medicare Data extracted based on the probabilistic linkage. Additionally, CCD data were delivered without SSN and HICN information. This resulted in 100% of CCDs having CMS Medicare Data extracted based on the probabilistic linkage and therefore the type II linkage error rate was not calculated.

Table 7. Algorithm Results by 2016 NHCS Record Source

Record Source	Cutoff	Eligible NHCS Patients	Total Links	Deterministic Matches	Non-Deterministic Links	Est Incorrect (Type I)	Est Not Found (Type II)
UB-04 Claims	0.85	3,294,026	771,231 (23.4%)	595,413 (77.2%)	175,818 (22.8%)	0.01%	0.17%
EHR Custom Extract	0.85	491,373	124,413 (25.3%)	111,041 (89.3%)	13,372 (10.7%)	<0.01%	0.05%
CCD	0.85	395,706	102,687 (26%)	0 (0%)	102,687 (100%)	0.12%	*

NOTES: Data are presented at patient level.
*Unable to estimate Type II linkage error for CCD records due to no SSN/HICN information on CCD records.

Appendix II: Descriptions of Medicare Data Files

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 & Utilization 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 segment** includes variables that indicate a Medicare FFS-enrolled beneficiary has received a service or treatment for selected chronic health conditions.[17] Additional information on each of the MBSF Segments may be found at

[17] Conditions Included in CCW: acquired hypothyroidism, acute myocardial infarction, Alzheimer's Disease, Alzheimer’s Disease & related disorders or senile dementia, anemia, asthma, atrial fibrillation, benign prostatic hyperplasia, cancer (colorectal), cancer (endometrial), cancer (female/male breast), cancer (lung), cancer (prostate), cataract, chronic kidney disease, chronic obstructive pulmonary disease (COPD), depression, diabetes, glaucoma,

https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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 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 significant claim-related event and date that may affect processing of payment by CMS; 4) a span code segment, identifying a significant 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.[18] Each segment is available as a separate file, but can be combined using the unique claim identification number (NCHS_CLM_ID) and unique NHCS Patient identifier (PATIENT_ID).

The Carrier and DME files share similar formats. Each file consists of a base claims segment, containing demographic information and diagnosis codes as well as billing and payment amounts associated with a non-institutionalized claim; and a line items segment that includes the specific billing and payment amounts for each line item included within the base claim; and a demonstrations 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) and unique NHCS 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

heart failure, hip / pelvic fracture, hyperlipidemia, hypertension, ischemic heart disease, osteoporosis, rheumatoid arthritis / osteoarthritis, stroke / transient ischemic attack

[18] CMS conducts various demonstration projects to test the impact of new methods of service delivery, coverage of new types of services, and new payment approaches: https://innovation.cms.gov/innovation-models (accessed August 18, 2020)

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. Additional information on the FFS IP File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the Encounter IP File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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 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. Additional information on the FFS SNF File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the Encounter SNF File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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 4.3.2) 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. Additional information on the FFS Carrier File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the Encounter Carrier File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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 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. Additional information on the FFS OP File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the Encounter OP File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the FFS DME File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the Encounter DME File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the FFS HHA File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the Encounter HHA File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the Hospice File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the MedPAR File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

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. Additional information on the PDE File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

5 Home Health Outcome and Assessment Information Set (OASIS)

The OASIS contains data items developed from patient assessments conducted to measure patient outcomes and to improve home health care. The OASIS assessments are required of all home health agencies certified to accept Medicare and Medicaid payments. OASIS data are collected for Medicare and Medicaid patients 18 years and older receiving skilled home health care services, with the exception of patients receiving services for pre- or postnatal conditions. Those receiving only personal care, homemaker, or chore services are excluded since these are not considered skilled services. OASIS data items include information on patient home environment and informal caregivers, functional status, psychosocial status, and health service utilization, including use of emergency services and hospital admission. Additional information on the OASIS File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).

6 Long-Term Care Minimum Data Set (MDS)

The Long-Term Care MDS is a health status screening and assessment tool used for all residents of long-term care nursing facilities certified to participate in Medicare or Medicaid. The assessment is also required for Medicare payment of SNF stays. MDS assessments are required for residents on admission to the nursing facility, periodically during the facility stay, and upon discharge. MDS data items include clinical status measures, psychological status, psychosocial functioning measures, physical functioning assessment, functional status, and end-of-life care decisions. Additional information on the MDS File may also be found at https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed August 18, 2020).