

The Linkage of the 2014 National Hospital
Care Survey to the 2014/2015 Centers for
Medicare & Medicaid Services
Medicare Enrollment, Claims and
Assessment Data: Methodology Overview
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 2014 National Hospital Care Survey to the 2014/2015 Centers for Medicare & Medicaid Services Medicare Enrollment, Claims, and Assessment Data: Methodology Overview and Analytic Considerations*, April 2024. Hyattsville, Maryland. Available at the following address: <https://www.cdc.gov/nchs/data-linkage/index.htm>

Contents

1 Introduction.....	5
2 Background on Linked Files	6
2.1 National Hospital Care Survey.....	6
2.2 Centers for Medicare & Medicaid Services, Medicare Data	6
3 Linkage Methodology	7
3.1 Linkage Eligibility Determination	7
3.2 Overview of Linkage	8
4 Analytic Considerations	10
4.1 Sampling Weights Are Currently Not Available	10
4.2 Hospital Linkage Eligibility.....	10
4.3 Patient_ID Details.....	10
4.4 Medicare Advantage	10
4.5 Cost Sharing	11
4.6 Medicare Payment and Conditions Data	11
4.7 Utilizing Administrative Race Data.....	12
4.8 On CMS Medicare MBSF Records with No Claims Data.....	12
4.9 Medicare Entitlement Variables	12
4.10 File Year Indicator	13
4.11 Analytic Considerations Specific to Medicare Fee-for-Service Claims Files.....	13
4.11.1 Carrier File.....	13
4.11.2 DME File.....	14
4.11.3 Hospice File.....	14
4.11.4 Outpatient File.....	15
4.11.5 Inpatient File.....	15
4.11.6 Medicare Provider Analysis and Review (MedPAR) File.....	15
4.12 Analytic Considerations Specific to the Medicare Part D Prescription Drug Event (PDE) File.....	16
5 Access to Data Files	17
5.0 Access to the Restricted-Use Linked NHCS – CMS Medicare Data	17
5.1 Combining the Linked NHCS-CMS Medicare Data to NHCS Analytic Files and Linked NDI data	17
Appendix I: Detailed Description of Linkage Methodology.....	19

1 Deterministic Linkage Using Unique Identifiers.....	19
2 Probabilistic Linkage.....	19
2.1 Blocking.....	19
2.2 Score Linked Pairs	20
2.3 Probability Modeling	23
3 Select Matches for Final File	23
Appendix II: Descriptions of Medicare Data Files.....	24
1 Master Beneficiary Summary File (MBSF).....	24
2 Standard Analytic Files (SAFs)	24
2.1 Fee-for-Service Inpatient (IP) Files.....	25
2.2 Fee-for-Service Skilled Nursing Facility (SNF) Files	25
2.3 Fee-for-Service Carrier Files.....	25
2.4 Fee-for-Service Outpatient (OP) Files	26
2.5 Fee-for-Service Durable Medicare Equipment (DME) Files.....	26
2.6 Fee-for-Service Home Health Agency (HHA) Files	26
2.7 Hospice File	26
3 Medicare Provider Analysis and Review (MedPAR) File	27
4 Medicare Part D Prescription Drug Event (PDE) File.....	27
5 Home Health Outcome and Assessment Information Set (OASIS)	27
6 Long-Term Care Minimum Data Set (MDS).....	28

1 Introduction

As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to provide statistical information that can be used to guide actions and policy to improve the health of the American people. In addition to collecting and disseminating the Nation's official vital statistics, NCHS conducts several population-based surveys and healthcare establishment surveys, including the National Hospital Care Survey (NHCS), <https://www.cdc.gov/nchs/nhcs/index.htm> (accessed August 6, 2019). The 2014 NHCS collected data from a sample of 581 hospitals, of which 95 participated by providing patient-level encounter records. For participating hospitals, these data cover all of their patient ambulatory care and inpatient visits occurring during the course of the year. The NHCS includes detailed information about hospital characteristics, patients' characteristics, and treatment. Even though NHCS is an establishment survey (i.e. hospitals are the sampling unit) it collects patient personally identifiable information (PII), which enable data linkages.

Through its data linkage program, NCHS has been able to expand the analytic utility of the data collected from NHCS by augmenting it with Medicare data collected from the Centers for Medicare & Medicaid Services (CMS). This report will describe the linkage of the 2014 NHCS to the 2014/2015 CMS Medicare enrollment, claims, and assessment data. Although NHCS is not currently nationally representative due to low response rates, $95/581=16\%$, linking NHCS with the CMS data allows for new analyses, such as examining comorbidities and utilization of non-inpatient related health care services.

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

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

2 Background on Linked Files

2.1 National Hospital Care Survey

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

From participating hospitals, NHCS collects data on all IP and ambulatory care visits occurring during the calendar year. Participating hospitals provide these data in the form of Uniform Bill (UB)-04 administrative claim records. Unlike its predecessor surveys, NHCS also collects PII (e.g., name, date of birth, and Social Security Number (SSN)), which allows for the linkage of episodes of care across hospital units and to other data sources, such as the CMS Medicare data. The linkage described in this document includes only IP and ED visits – other, non-ED OPD visits have been excluded.

2.2 Centers for Medicare & Medicaid Services, Medicare Data

The 2014 NHCS has been linked to CMS Medicare enrollment, claims, and assessment data from 2014–2015.

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 2014 and 2015, approximately 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 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 enrollment status, health care utilization, and expenditures.

The CMS Medicare Master Beneficiary Summary File (MBSF) is an annual file containing demographic and enrollment information about beneficiaries enrolled in Medicare during each

calendar year. The CMS Medicare MBSF includes information on beneficiary demographic characteristics, reason for Medicare entitlement, and program enrollment type (Original Medicare vs. Medicare Advantage (MA)).

The 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 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

Linkages were only possible for NHCS patient records that had certain minimum levels of PII available. In order for a record to be considered linkage eligible, the record must have valid date of birth (month, day, and year)¹ and name (first, middle, and last)² information present. For example, if the PII on the NHCS claims records had a full name and only the year of birth, they were deemed as being ineligible for linkage, as there would be a considerable likelihood of not being able to find a CMS match even if one exists.

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

¹ A date of birth is considered to be usable if at least two of the three date parts are valid date values.

² A name is considered to be usable if at least two of these three criteria is met: first name has two or more characters, middle name has one or more characters, and last name has two or more characters.

3.2 Overview of Linkage

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

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

The linkage between the 2014 NHCS records and the CMS EDB was based on both deterministic and probabilistic approaches. The probabilistic approach performs weighting and link adjudication as described in the Fellegi-Sunter paradigm method.³ Following these methods, a selection process was implemented with the goal of selecting pairs believed to represent the same individual between the data sources. Table 1 highlights the linkage results. In sum, the linkage steps are the following (to be explained in further detail in the appendix):

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

For each NHCS record that was linked, CMS extracted data records from its 2014 and 2015 SAFs and provided them to NCHS.

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

	Sample Size			Percent Linked	
	Total Sample	Eligible for Linkage ²	Linked to 2014-2015 Medicare Administrative Data ³	Total Sample ⁴	Eligible Sample ⁵
Age¹					
<65	2,946,281	2,685,538	234,527	8.0%	8.7%
>=65	610,784	550,221	538,451	88.2%	97.9%
Total	3,557,065	3,235,759	772,978	21.7%	23.9%

³ Fellegi, I. P., and Sunter, A B. (1969), "A Theory for Record Linkage," JASA 40 1183-1210.

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

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

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

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

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

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

4 Analytic Considerations

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

4.1 Sampling Weights Are Currently Not Available

Currently, there are no sampling weights available for the 2014 NHCS data. This section will be updated if sampling weights are made available in the future. Because the hospital level sampling conducted for NHCS was not conducted on an equal probability basis, unweighted estimates will be biased to be more similar to those from hospitals selected with higher sampling probability. Similarly, there will be bias towards types of hospitals responding at higher rates. These biases will be more of a concern if estimates vary strongly by factors correlated with sampling and response rates. One way to mitigate these biases is to calculate estimates in the framework of regression modeling that controls for hospital characteristics. This would be done by including hospital characteristics (region, ownership type, and size) as well as patient characteristics (age and sex) among the predictor variables in the model definition. Statistical testing can then be conducted on parameter estimates associated with these characteristics.

4.2 Hospital Linkage Eligibility

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

4.3 Patient_ID Details

PATIENT_ID is a de-identified ID that is intended to be unique for each individual receiving IP, ED, or OPD services at a participating hospital. However, since the de-duplication of patient records required to generate this ID depends on sometimes incomplete or erroneous data, there may be instances where the same individual is represented by more than one **PATIENT_ID**. This happens infrequently and should not greatly impact analyses.⁴

4.4 Medicare Advantage

CMS generally does not receive fee-for-service claims for Medicare beneficiaries who are enrolled in Medicare Advantage (including private fee-for-service plans paid on a capitation basis). Medicare Advantage plans are also referred to as Medicare Part C and include Health

⁴ For more information of Patient_ID generation, see Technical Notes on page 14: <https://www.cdc.gov/nchs/data/nhsr/nhsr097.pdf> (accessed August 6, 2019)

Maintenance Organizations (HMOs), Preferred Provider Organizations (PPOs), Private Fee-for-Service (PFFS) Plans, Special Needs Plans, and Medicare Medical Savings Account Plans. During the time covered by the linked data files, Medicare Advantage enrollment reached 31% of total Medicare beneficiaries.

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

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

4.5 Cost Sharing

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

4.6 Medicare Payment and Conditions Data

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

The CMS Medicare MBSF Chronic Conditions segment flags each Medicare beneficiary for the presence of one of 27 specific chronic conditions. Additional information about the methodology used to assign chronic condition flags to Medicare beneficiaries is available at <https://www.cwdata.org/web/guest/condition-categories> (accessed August 6, 2019). CMS

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

4.7 Utilizing Administrative Race Data

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

4.8 On CMS Medicare MBSF Records with No Claims Data

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

4.9 Medicare Entitlement Variables

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

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

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

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

4.10 File Year Indicator

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

4.11 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.

4.11.1 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.

4.11.2 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.

4.11.3 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.

4.11.4 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.

4.11.5 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.

4.11.6 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.

4.11.6 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 [2014 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.

4.12 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

To ensure confidentiality of health data, NCHS provides safeguards including the removal of all personal identifiers from analytic files. Additionally, the files containing these data are only made available in secure facilities for approved research projects. Researchers who want to obtain the linked 2014 NCHS- 2014/2015 CMS Medicare 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 6, 2019).

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

NHCS is an establishment survey where the respondents are individual hospitals rather than their patients. Typically this type of survey restricts analyses to the sample unit-level, but

because NHCS collects hospital encounter-level records, encounter-level analysis is also possible. For each patient with either an IP discharge or ED visit, results of the person-level linkage to the CMS Medicare data are available in the linked 2014 NHCS-2014/2015 CMS Medicare data.

To perform encounter-level analysis, the linked 2014 NHCS-2014/2015 CMS Medicare data files can be used in conjunction with 2014 NHCS analytic files⁵ and the linked 2014 NHCS-2014/2015 NDI mortality file, which are also available through the NCHS RDC. The linked mortality file includes Patient_ID, date of birth, date of death, and cause of death information, while the analytic files include analytically-pertinent hospital-level details (such as bed size and geographic region) and episode-level details (patient demographics, diagnoses, procedures, admission and discharge dates).

To integrate the NHCS analytic and the linked 2014 NHCS-2014/2015 mortality file into the linked 2014 NHCS- 2014/2015 CMS Medicare data, joins should be made on the common field, **PATIENT_ID**. Additionally, **PATIENT_ID** allows linkage of multiple visits for the same patients within or across hospital settings (IP or ED).

⁵ Find more information about the NHCS analytic file: <https://www.cdc.gov/rdc/b1datatype/dt1224h.htm> (accessed August 6, 2019)

Appendix I: Detailed Description of Linkage Methodology

1 Deterministic Linkage Using Unique Identifiers

The first step in the linkage process is to attempt a deterministic linkage for all eligible NHCS records that were submitted with a valid format SSN or HICN. In some cases, the SSN field does not hold a validly formatted value and a possible replacement value using the HICN is instead used.⁶ Using HICN to derive a SSN when no valid SSN was available resulted in a 13.35% increase in IP ID's with a valid SSN.

The deterministic linkages were validated by comparing first name, middle initial, last name, month of birth, day of birth, year of birth, zip code of residence, and state of residence identifiers in order to ensure that the records were a valid match. If the ratio of matching identifiers to non-missing identifiers is greater than 50%, the linked pair is retained as a deterministic match. The collection of records resulting from the deterministic match is referred to as the 'truth deck.'

2 Probabilistic Linkage

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

2.1 Blocking

Blocking is a key step in record linkage. It identifies potential candidate pairs without comparing every single pair in the Cartesian product. According to Christen, blocking or indexing, “splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key).”⁷ Rules can be used to define the blocking criteria however, for this linkage, instead of rules, we used the data to help create a set of blocks that would efficiently join the datasets together. By using the data to create the efficient block set, we ideally reduce the number of false positive links while retaining a high percentage of true positive links. For the purpose of this linkage, the 'truth deck' was used as the training dataset. When the data are used in this manner, it is commonly referred to as a machine learning algorithm. For more detailed information on the method that was used please refer to “Learning Blocking Schemes for Record Linkage.”⁸

⁶ Only the HICN's where the individual was listed as the primary beneficiary were used as replacements for SSNs

⁷ Christen, Peter. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Berlin Heidelberg: Springer-Verlag, 2012. <http://www.springer.com/us/book/9783642311635> (Accessed August 6, 2019).

⁸ 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/d721845dd876c769c1fd2d967c04f3a6eaaa.pdf> (Accessed August 6, 2019).

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

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

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Block 7
<ul style="list-style-type: none"> • Last Name • Full DOB • Zip Code 	<ul style="list-style-type: none"> • Last Name • Full DOB • Sex 	<ul style="list-style-type: none"> • First Name • Full DOB • Sex 	<ul style="list-style-type: none"> • State ID • Full DOB • Sex 	<ul style="list-style-type: none"> • Last Name • Month of Birth • State • Zip Code 	<ul style="list-style-type: none"> • Last Name • First Name • State • Sex 	<ul style="list-style-type: none"> • Year of Birth • Sex • State • Zip Code
Score 1	Score 2	Score 3	Score 4	Score 5	Score 6	Score 7
<ul style="list-style-type: none"> • First Name • Middle Name • Sex 	<ul style="list-style-type: none"> • First Name • Middle Name • State • Zip Code 	<ul style="list-style-type: none"> • Middle Name • Last Name • State • Zip Code 	<ul style="list-style-type: none"> • First Name • Middle Name • Last Name • Zip Code 	<ul style="list-style-type: none"> • First Name • Middle Name • Year of Birth • Day of Birth • Sex 	<ul style="list-style-type: none"> • Middle Name • Month of Birth • Year of Birth • Day of Birth • Zip Code 	<ul style="list-style-type: none"> • First Name • Middle Name • Last Name • Month of Birth • Day of Birth

2.2 Score Linked Pairs

Next, each linked pair is scored using an approach based on the Fellegi-Sunter paradigm. The Fellegi-Sunter paradigm specifies the functional relationship between agreement probabilities and agreement/non-agreement weights for each identifier used in the linkage process. The scores – pair weights – calculated in this step are used in a probability model (explained in [Section 2.3](#)), which allows the linkage algorithm to select final pairs to include in the matched file. The scoring process follows the following order:

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

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

- First Name or First Initial (when applicable)

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

2.2.1 Calculate M- and U- Probabilities

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

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

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

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

Like the M-probabilities, only variables outside of the blocking factor have a U-probability calculated. The U-probabilities are calculated using the frequency of each value with in a variable in the full set of NHCS submission records. For example, the U-probability of a state with many patients would be around 0.06 (6.0%) but for a state with less patients the U-probability would only be about 0.0003 (0.03%) because records from individuals residing in that state are less common in the data file. Similar to the M-probabilities, first and last name are not calculated in the same manner as the rest of the identifiers. The calculation of the U-probabilities for first and last name are discussed in much greater detail in the following section.

2.2.2 M and U Probabilities for First and Last Names

Similar to the M-probabilities, Jaro-Winkler levels (85, 90, 95, and 100) are also calculated for use in the U-probability section. The manner of their creation is identical to the process

described above. For several reasons, the first and last name U probabilities were computed differently than for the remaining comparison variables. Because of the many possible values for first and last name, it was impractical to compute U- probabilities specific to each blocking factor. Instead, we computed U-probabilities using all records in the NHCS submission file and a simple random sample of 1% of the CMS Medicare EDB submission file.

Complete name tallies (separately, for first and last names) were then produced for the NHCS submission file. For each level of name on the file, we randomly selected 100,000 names from the CMS Medicare EDB submission file 1% sample to compare to it. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85, and for each NHCS name, we tallied the number of the 100,000 randomly selected CMS Medicare EDB names that agreed at that level.^{9,10}

2.2.3 Calculate Agreement and Non-Agreement Weights

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

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

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

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

2.2.4 Calculate Pair Weight Scores

The next step is to calculate pair weights, which are used in the probability model. The pair weights are calculated differently for each linked record pair, but follow a same general process:

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

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

⁹ 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.

¹⁰ Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association. 1990. 354-9.

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

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

2.3 Probability Modeling

A probability model, developed from a logistic regression analysis, estimates the likely validity that each pair is a match. Those pair-specific probabilities allows the linkage algorithm (as described in [Section 2.2](#)) to first determine which pair to select when multiple pairs are available for a given patient ID, and then determine whether the pair's likely validity is great enough to keep in the final set of accepted matches.

The logistic model was estimated on possible matched pairs that had a valid and formatted SSN on both the NHCS patient record and the CMS enrollment record. The response variable used in the regression analysis is the agreement on SSN, created as a categorical variable (1 – SSN values agree, 0 – SSN values do not agree).¹¹ The regression model estimates the likely validity that each pair is a match given its pair weight. Since multiple pairs may exist for the same patient, the model allows the linkage algorithm to calculate the probability of validity for each possible pair, which can then be used to identify the pair with the highest probability of being a valid match.

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

3 Select Matches for Final File

Up to this point, the linkage has identified possible matches through both the deterministic linkage and the probabilistic linkage. These identified matches all have a probability value assigned that measures their probability of being a valid match. The deterministic matches were automatically assigned a probability value of (1), while the probabilistic links were assigned a probability of validity¹² using the logistic model.

The penultimate step is to assign a probability threshold that a pair is a valid match. The probability threshold for the linked 2014 NHCS – 2014/2015 CMS Medicare Data is set to 0.85. If the best possible match has an estimated validity less than the threshold, then the linkage algorithm will not accept it into the final matched file.

Last, the linkage algorithm selects only one pair per patient on the NHCS file – of those pairs that met the probability thresholds just discussed – to include in the final matched file. If there is only one possible pair for a given patient above the relevant threshold, then that pair is included

¹¹ The linkage classifies it as an agreement if five or more of the nine SSN positions have the same digit on the SSN values being compared.

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

in the final file. If there is more than one possible matched pair for a given patient above the relevant threshold, then the possible matched pair with the highest probability of being a valid match is selected. If a tie remains at this point then the record with the better matching information is selected. If all information is matching the same, one record is selected at random. Note – if one of these possible pairs was created during the deterministic linkage, then this pair will always be selected because it is assigned a probability of one.

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.¹³ Additional information on each of the MBSF Segments may be found at <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 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

¹³ 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, heart failure, hip / pelvic fracture, hyperlipidemia, hypertension, ischemic heart disease, osteoporosis, rheumatoid arthritis / osteoarthritis, stroke / transient ischemic attack

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.¹⁴ 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 demonstration code segment. The base claim, line item, and demonstration code segments are available as separate files but can be combined using the unique claim identification number (NCHS_CLM_ID) and unique NHCS Patient identifier (PATIENT_ID).

2.1 Fee-for-Service Inpatient (IP) Files

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

2.2 Fee-for-Service Skilled Nursing Facility (SNF) Files

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

2.3 Fee-for-Service Carrier Files

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

¹⁴ 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)

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).

2.4 Fee-for-Service Outpatient (OP) Files

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

2.5 Fee-for-Service Durable Medicare Equipment (DME) Files

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).

2.6 Fee-for-Service Home Health Agency (HHA) Files

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).

2.7 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).