The Linkage of National Center for Health Statistics Survey Data to Medicare Enrollment, Claims/Encounters and Assessment Data (2014-2018):

Linkage Methodology and Analytic Considerations

Data Release Date: April 2024
Document Version Date: April 23, 2024

Suggested citation:
National Center for Health Statistics, Division of Analysis and Epidemiology. The Linkage of National Center for Health Statistics Survey Data to Medicare Enrollment, Claims/Encounters and Assessment Data (2014-2018): Linkage Methodology and Analytic Considerations. April 2024. Hyattsville, Maryland.

Available at the following address:
https://www.cdc.gov/nchs/data-linkage/medicare-methods.htm

Table of Contents

List of Acronyms

AMA, American Medical Association

BIC, Beneficiary Identification Code

CCD, Consolidated Clinic Document

CMS, Center for Medicare & Medicaid Services

COPD, Chronic Obstructive Pulmonary Disease

CPT-4, Current Procedural Terminology, 4th Edition

DME, durable medical equipment

DMERC, durable medical equipment regional carrier

DOB, date of birth

DSH, disproportionate share

EDB, enrollment database

ESRD, end-stage renal disease

ERB, ethics review board

FFS, fee-for-service

GME, graduate medical education

HCPCS, Healthcare Common Procedure Coding System

HHA, home health agency

HICN, Health Insurance Claim Number

HMO, health maintenance organization

HUD, Department of Housing and Urban Development

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

IME, indirect medical education

IP, inpatient

MA, Medicare Advantage

MAC, Medicare Administrative Contractor

MAO, Medicare Advantage Organization

MA-PD, Medicare Advantage Prescription Drug Plan

MBSF, Master Beneficiary Summary File

MDS, Minimum Data Set

MedPAR, Medicare Provider Analysis and Review File

NCHS, National Center for Health Statistics

NDI, National Death Index

NHANES, Continuous National Health and Nutrition Examination Survey

NHANES III, Third National Health and Nutrition Examination Survey

NHIS, National Health Interview Survey

NNHS, National Nursing Home Survey

OASIS, Outcome and Assessment Information Set

OP, outpatient

OTC, over-the-counter

PDE, prescription drug event

PDP, prescription drug plan

PII, personally identifiable information

PPO, preferred provider organization

RDC, Research Data Center

RDS, retiree drug subsidies

ResDAC, Research Data Assistance Center

SAF, standard analytic file

SNF, skilled nursing facility

SNP, special needs plan

SSN, Social Security number

VA, Veteran Affairs

VRDC, Virtual Research Data Center

The Linkage of National Center for Health Statistics Surveys to Medicare Enrollment, Claims/Encounters and Assessment Data (2014-2018): Linkage Methodology and Analytic Considerations

1 Introduction

As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to provide statistical information that can be used to guide actions and policy to improve the health of the American people. As part of its ongoing efforts to fulfill this mission, NCHS conducts several population-based and establishment surveys that provide rich cross-sectional information on risk factors such as smoking, height and weight, health status, and socio-economic circumstances. Although the survey data collected provide information on a wide-range of health-related topics, they often lack information on longitudinal outcomes. Through its data linkage program, NCHS has been able to enhance the survey data it collects by supplementing survey information with information from health-related administrative data sources. The linkage of survey and administrative data provide the unique opportunity to study changes in health status, health care utilization and expenditures in specialized populations, such as elderly people and people with disabilities.

Under an interagency agreement between NCHS and the Centers for Medicare & Medicaid Services (CMS) several NCHS surveys have been linked to 2014-2018 Medicare enrollment, Medicare Part A and B beneficiary fee-for-service (FFS) health care claims, Medicare Part D prescription drug events, and home health and long term care patient assessments, and 2016–2018 Medicare Advantage (MA) beneficiary encounter data.

The resulting linked data files provide the opportunity to examine the administrative data during the year the survey was conducted, in years following the survey, as well as the years prior to the survey for some NCHS survey participants. The linked NCHS-Medicare files, in particular, combine health and socio-demographic information from the surveys with enrollment, claims/encounters and assessment information from the Medicare program, resulting in unique population-based information that can be used for an array of epidemiologic and health services research and to support evidence-based policy evaluation.

This report describes the most recent linkage conducted between selected NCHS surveys and CMS administrative records. A brief overview of the data sources, the methods used for linkage, descriptions of the resulting linked data files, and analytic guidance are provided in this report. More information about the previous linkages of NCHS survey and Medicare data is available in the following reports:

- *The Linkage of National Center for Health Statistics Surveys to Medicare Enrollment and Claims Data (1999-2013)- Methodology and Analytic Considerations* (https://www.cdc.gov/nchs/data-linkage/cms/nchs_medicare_linkage_methodology_and_analytic_considerations.pdf (published September 2017))

- *Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare & Medicaid Services* (http://www.cdc.gov/nchs/data/series/sr_01/sr01_058.pdf (published September 2015))

2 Data Sources

2.1 National Center for Health Statistics, Survey Data

NCHS has recently linked the following surveys to 2014-2018 Medicare enrollment, FFS claims, prescription drug events, patient assessments and 2016–2018 MA encounter data[1]:

- 1994-2018 National Health Interview Survey (NHIS)
- 1999-2018 Continuous National Health and Nutrition Examination Survey (NHANES)
- Third National Health and Nutrition Examination Survey (NHANES III)
- 2004 National Nursing Home Survey (NNHS)

More detailed information regarding specific CMS Medicare data file availability and the linked data time span for these NCHS surveys is provided in Section 4.1.8.

Additionally, NCHS previously linked 1999-2013 CMS Medicare enrollment and FFS claims data for the following NCHS surveys:

- 1994-2013 NHIS
- Second Longitudinal Study of Aging (LSOA II)
- 1999-2012 Continuous NHANES
- NHANES III
- NHANES I Epidemiologic Follow-Up Study (NHEFS)
- 2004 NNHS

More information regarding the 1999-2013 NCHS-CMS linked data files can be found in the reports describing the previous linkages referenced in Section 1.

A brief description of the NCHS surveys included in the updated CMS Medicare linkages follows.

NHIS is a nationally representative, cross-sectional household interview survey that serves as an important source of information on the health of the civilian, noninstitutionalized population of the United States. It is a multistage sample survey with primary sampling units of counties or adjacent counties, secondary sampling units of clusters of houses, tertiary sampling units of households, and finally, persons within households. It has been conducted continuously since 1957 and the content of the survey is periodically updated. NHIS has been used as the sampling frame for other NCHS surveys focusing on specialized populations, including **LSOA II**. Prior to 2007, NHIS traditionally collected full 9-digit Social Security Numbers (SSN) from survey participants. However, in attempt to address respondents' increasing refusal to provide SSN and consent for linkage, NHIS began, in 2007, to collect only the last 4 digits of SSN and added an explicit question about linkage for those who refused to provide SSN. The implications of this procedural change on data linkage activities are discussed later in this report. NHIS

[1] The initial data release in April 2021 did not include 2014–2015 FFS claims, prescription drug events, and patient assessments or 2017–2018 MA encounters. The updated data release in April 2024 added those files for most linked surveys, with the exception of NHIS 1994–1998 and NNHS 2004.

implemented a content and structure redesign in 2019. For detailed information on the NHIS's contents and methods, refer to the NHIS website, http://www.cdc.gov/nchs/nhis.htm (accessed October 27, 2020).

NHANES is a continuous, nationally representative survey consisting of about 5,000 persons from 15 different counties each year. For a variety of reasons, including disclosure issues, the NHANES data are released on public-use data files in two-year increments. The survey includes a standardized physical examination, laboratory tests, and questionnaires that cover various health-related topics. NHANES includes an interview in the household followed by an examination in a mobile examination center (MEC). NHANES is a nationally representative, cross-sectional sample of the U.S. civilian, noninstitutionalized population that is selected using a complex, multistage probability design.

Prior to becoming a continuous survey in 1999, NHANES was conducted periodically, with the last periodic survey, **NHANES III**, conducted between 1988 and 1994. NHANES III was designed to provide national estimates of health and nutritional status of the civilian, non-institutionalized population of the United States aged 2 months and older. Similar to the continuous survey, NHANES III included a standardized physical examination, laboratory tests, and questionnaires that covered various health-related topics.

For detailed information about the Continuous NHANES and NHANES III contents and methods, refer to the NHANES website, https://www.cdc.gov/nchs/nhanes/index.htm (accessed October 27, 2020).

NNHS provides information on nursing homes from two perspectives- that of the provider of services and that of the recipient of care. Data for the surveys were obtained through personal interviews with facility administrators and designated staff who used administrative records to answer questions about the facilities, staff, services and programs, and medical records to answer questions about the residents. NNHS was first conducted in 1973-1974 and repeated in 1977, 1985, 1995, 1997, 1999, and most recently in 2004. Only the 2004 survey was included in this CMS Medicare linkage. For more information on the NNHS content and methods, refer to the NNHS website, http://www.cdc.gov/nchs/nnhs.htm (accessed October 27, 2020).

2.2 Centers for Medicare & Medicaid Services, Medicare Data

These NCHS survey data have been linked to CMS Medicare data from 2014-2018.

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). During 2014-2018, approximately two-thirds of persons enrolled in Medicare, known as Medicare beneficiaries, were enrolled in traditional Medicare, also known as Medicare FFS. Nearly all Medicare FFS beneficiaries receive Part A hospital insurance benefits, which help cover inpatient (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, outpatient (OP) care, durable medical equipment (DME), and some home health care services.

During 2014-2018, 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 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 I: Descriptions of Medicare Data Files.

3 Linkage of NCHS Survey Data with 2014-2018 Medicare Records

3.1 Linkage Eligibility

The linkage of NCHS survey participants to their Medicare enrollment, claims/encounters and assessment data was conducted under an interagency agreement between NCHS and CMS. The linkage was performed by NCHS in the CMS Virtual Research Data Center (VRDC) and is not the responsibility of researchers using the data. Approval for the linkage was provided by NCHS's Research Ethics Review Board (ERB)[2] and the linkage was performed only for eligible NCHS survey participants. Only NCHS survey participants who have provided consent as well as the necessary personally identifiable information (PII), such as date of birth and full or partial SSN or Medicare Health Insurance Claim Number (HICN), are considered linkage-eligible. Linkage-eligibility refers to the potential ability to link data from an NCHS survey participant to administrative data. Due to variability of questions across NCHS surveys, changes to PII collection procedures by the surveys over time, and changes in who is asked specific questions, criteria for NCHS-CMS Medicare linkage eligibility vary by survey and year.

For many of the surveys initiated prior to and during 2007, including 1994-2006 NHIS, 1999-2008 NHANES, NHANES III, and 2004 NNHS, a refusal by the survey participant to provide an SSN or HICN was considered an implicit refusal for data linkage. However, NCHS began to notice an increase in the refusal rate for providing SSN and HICN, particularly for NHIS, which reduced the number of survey participants eligible for linkage (1). In attempt to address declining linkage eligibility rates, NCHS introduced new procedures for obtaining consent for linkage from survey participants. Research was also conducted to assess the accuracy of matching data from NHIS to the National Death Index (NDI) using partial SSN and other PII (2). The research assessed algorithms using the last four and last six digits of SSN. The results were favorable and provided sufficient data to support changes in how NHIS collected SSN and HICN for linkage (3). Beginning in 2007, NHIS started requesting only the last four digits of SSN and HICN (plus a letter) instead of the complete number for both identifiers. In addition, a short introduction before asking for SSN was added and participants who refused to provide SSN or HICN were asked for their explicit permission to link to administrative records without SSN or HICN. Also, at this time, the NCHS ERB determined that for 2007 NHIS and all subsequent years, only primary respondents (sample adult and sample child) would be eligible for linkage to administrative records.

The informed consent procedures changed for the continuous NHANES as well. NHANES continued to collect full nine-digit SSN and complete HICN through the 2017-2018 survey cycle. However, beginning with the 2009-2010 NHANES, participants were explicitly asked for consent to be included in data linkage activities during the informed consent process prior to the interview. Only participants who provided an affirmative response to the linkage question were

[2] The NCHS Research Ethics Review Board (ERB), also known as an Institutional Review Board or IRB, is an administrative body of scientists and non-scientists that is established to protect the rights and welfare of human research subjects.

considered linkage eligible. In addition, starting in 2017-2018, survey participants who consented to linkage but who refused to provide their full nine-digit SSN and complete HICN were given the option to provide only the last four digits of either identification number.

3.2 Child Survey Participants

NCHS survey participants under 18 years of age at the time of the survey are considered linkage-eligible if the linkage eligibility criteria described above are met and consent is provided by their parent or guardian. However, the consent provided by the parent or guardian does not apply once the child survey participant becomes a legal adult, and there is no opportunity for NCHS to obtain consent to link the child participant's survey data to administrative data based on their adult experiences. As a result, in accordance with NCHS ERB guidance, NCHS only includes administrative data that were generated for program participation, claims and other events that occurred prior to the participant's 18th birthday on the linked data files provided to researchers. Since the majority of Medicare beneficiaries are age 65 and older, the ERB guidance pertains to less than 1% of the linked survey participants in the Medicare linkage.

3.3 CMS Virtual Research Data Center

The linkage of NCHS survey data with Medicare administrative data was performed by authorized NCHS staff within the CMS VRDC. The CMS VRDC is a virtual research environment that allows approved researchers to access Medicare and Medicaid program data from their own personal workstations. VRDC users are granted direct access to approved CMS data files and are able to conduct analyses for research projects within a secure environment. VRDC users have the ability to download aggregated reports and results to their personal workstations, following disclosure review. More information about the CMS VRDC can be found at https://www.resdac.org/cms-virtual-research-data-center-vrdc (accessed October 27, 2020).

3.4 Linkage Methodology

This section outlines steps that were used to link the NCHS survey data with 2014-2018 CMS Medicare Enrollment Database (EDB). The CMS EDB is the database of record for Medicare Beneficiary enrollment information and includes Medicare beneficiary identification information. For more detailed information on linkage methodology (see Appendix II: Detailed Description of Linkage Methodology).

Records for NCHS survey participants were linked to records in the CMS EDB using the following identifiers: SSN (9 digits or last 4 digits, depending on the survey and year of the survey), HICN (complete identifier or last 4 digits plus a letter, depending on the survey and year of the survey), 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 NCHS survey participant 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 (4). Following this, a selection process was implemented with the goal of selecting pairs believed to match (i.e., representing the same individual between the data sources). The following three steps (explained in further detail in

Appendix II: Detailed Description of Linkage Methodology) were applied to determined matched records:

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

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

3.5 Linkage Rates

The linkage rates for adults for the 2014-2018 NCHS-CMS Medicare linkage are provided in tables which can be accessed at this location: https://www.cdc.gov/nchs/data/datalinkage/cms_medicare_linked_data_match_rate_tables_1.pdf. For each linked NCHS survey, the tables present the total survey sample size, the sample size eligible for the Medicare linkage, the number of eligible survey participants linked to the CMS EDB and the linkage rate for both the total survey sample and the linkage eligible survey sample. The eligible survey sample includes only survey participants who were considered eligible for linkage as previously described. Linkage eligibility did not account for vital status, and participants may have been eligible for linkage even if they had died prior to the administrative data years.

Medicare has age-based entitlement at age 65. Therefore, the linkage rates for each survey were examined overall and by two age groups – 18-64 years and 65 years and older. Age was defined as the survey participant's age at interview. For the earliest linked surveys, the match rate is lower among adults aged 65 and over at interview compared with adults aged 18–64 at interview. This is because some adults aged 18–64 at interview have aged into Medicare eligibility before the end of the administrative data years, while some adults aged 65 and over at interview have died prior to the administrative data years.

4 Analytic Considerations when using the Linked NCHS-CMS Medicare Files

This section summarizes some key analytic issues for users of the linked NCHS survey data and CMS administrative records. It is not an exhaustive list of the analytic issues that researchers may encounter while using the Linked NCHS-CMS Medicare Data Files. 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 NCHS-CMS Medicare linked data are encouraged to visit the Research Data Assistance Center (ResDAC) website http://www.resdac.org/ (accessed October 27, 2020) for more information on Medicare data and their analytic considerations.

4.1 General Analytic Guidance for Data Users

4.1.1 Access to the Restricted-Use Linked NCHS-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 NCHS - 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 October 27, 2020).

4.1.2 Linked NCHS-CMS Medicare Data Feasibility Files

NCHS has created public-use Linked NCHS-CMS Medicare Data Feasibility Files to assist researchers who are considering submitting an RDC application to analyze the linked NCHS-CMS Medicare data. Feasibility Files are currently available for the NCHS surveys linked with 1999-2013 Medicare data and 2014-2018 Medicare data at https://www.cdc.gov/nchs/data-linkage/medicare-feasibility.htm (accessed May 24, 2021).
The feasibility files contain information on survey participants' eligibility status for linkage, match status, and variables indicating which specific CMS Medicare data files are available for each successfully linked survey participant for the datasets in the initial data release in April 2021. The feasibility files have not been updated with information about whether linked survey participants have data in the additional files in the April 2024 release (i.e. 2014–2015 claims files and 2017–2018 MA encounter files). The feasibility files remain available for researcher use, and the updated data release has no impact on the published information.

4.1.3 Merging Linked NCHS-CMS Medicare Data with NCHS Survey Data

To perform person-level analysis, the restricted-use Linked NCHS-CMS Medicare Data Analytic Files and public-use Linked NCHS-CMS Medicare Data Feasibility Files can be used in conjunction with the NCHS collected survey data (described above in Section 2.1). A unique survey participant identification variable is available on each file that allows analysts to merge survey data for survey participants with their information from the NCHS-CMS Medicare Linked Data files. The unique survey identifiers are survey-specific and may be constructed differently across survey years. Please refer to Appendix III: Merging Linked NCHS-CMS Medicare Files with NCHS Survey Data for guidance on identifying and constructing (if necessary) the appropriate identification variable for merging survey data and the NCHS-CMS Medicare Linked Data files.

For more information on the variables required to link information across NCHS-CMS Medicare Linked files, please see Appendix I, Section 2.

4.1.4 Variables to Request in RDC Proposals

To create analytic files for use in the RDC, a researcher provides a file containing the variables from the public-use NCHS survey data to RDC for merging with the requested restricted variables from NCHS surveys and for use with the variables from the CMS linked data files. The restricted variables from NCHS surveys and the exact variables from the CMS linked data files that the researcher will use also need to be specifically requested as part of a researcher's application to RDC. Staff in the RDC verify the full list of variables (restricted and public-use) and check for potential disclosure risk.

Although the complete list of variables used for specific analyses differs, the following variables from NCHS surveys should be considered for inclusion:

- Geography— Geography information is available on the administrative data for linked participants. However, there may be differences in the information available from the survey and administrative data. It is recommended that users who require information on geography request this information from the NCHS survey.

- Linked mortality data for NCHS surveys—Each of the NCHS surveys that have been linked to the Medicare data have also been linked to death information obtained from the NDI. The linked NDI mortality files provide date and cause of death for each survey participant who has died. Researchers interested in analyzing linked mortality data with linked CMS data must specifically request the desired mortality variables in their RDC proposal. More information about the NCHS-NDI linked mortality files can be found at https://www.cdc.gov/nchs/data-linkage/mortality.htm (accessed October 27, 2020).

- NHANES month and year of examination and interview—NHANES is released in 2-year cycles. The exact year (and month) of a survey participant's interview and examination are not provided on public-use files. However, many researchers will want to know the time elapsed between a given year (or even month) of the Medicare data and the NHANES interview or examination. The variables that indicate the month and year of NHANES interview or examination must be requested specifically.

It is recommended that researchers request the following variables, available from the public-use NCHS survey files, for inclusion in analytic files:

- Sample weights and design variables—these variables are needed to account for the complex design of the NCHS surveys. The names of the weights and design variables differ depending on which NCHS survey is being used. These can be identified using the documentation for each NCHS survey. As discussed below, NCHS recommends adjusting the sample weights to account for linkage eligibility bias.

- Demographic information about survey participants from the NCHS survey— For variables such as race and ethnicity, NCHS demographic information is self- or family respondent-reported and, thus, may be more accurate than demographic data provided in the Medicare files. Therefore, when possible, the NCHS data should be used for demographic variables.

Note: All RDC applications to analyze linked NCHS-CMS data should include requests to analyze the Master Beneficiary Summary File (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.

4.1.5 Sample Weights

The sample weights provided in NCHS population health survey data files adjust for oversampling of specific subgroups and differential nonresponse and are post-stratified to annual population totals for specific population domains to provide nationally representative estimates. The properties of these weights for linked data files with incomplete linkage, due to ineligibility for linkage, are unknown. In addition, methods for using the survey weights for some longitudinal analyses require further research. Because this is an important and complex methodological topic, ongoing work is being done at NCHS and elsewhere to examine the use of survey weights for linked data in multiple ways.

One approach is to analyze linked data files using adjusted sample weights. The sample weights available on NCHS population health survey data files can be adjusted for linkage eligibility (nonresponse), using standard weighting domains to reproduce population counts within these domains: sex, age, and race and ethnicity subgroups. These counts are called “control totals” and are estimated from the full survey sample.

A model-based calibration approach developed within the SUDAAN software package (Procedure WTADJUST or WTADJX) allows auxiliary information to be used to adjust the sample weights for nonresponse. This approach is recommended for adjusting sample weights for the linked files. Because inferences may depend on the approach used to develop weights, within SUDAAN’s WTADJUST or using a different calibration approach, researchers should seek assistance from a statistician for guidance on their particular project. Other approaches or software can be used. NCHS continues to investigate alternate approaches for addressing issues related to missing data, including the use of multiple imputation techniques. More detailed information on adjusting sample weights for linkage eligibility using SUDAAN can be found in Appendix III of *Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare & Medicaid Services* (5). To calculate the adjusted weights for linkage eligibility it is suggested that researchers use the variables CMS_MEDICARE_MATCH (for the 1999–2013 linkage) or MEDICARE_MATCH_1418 (for the 2014–2018 linkage) from the public-use feasibility files.

4.1.6 Temporal Alignment of Survey and Administrative Data

NCHS surveys have been linked to multiple years of Medicare administrative data. Depending on the survey year, Medicare data may be available for survey participants at the time of the survey, as well as before or after the survey period. Several factors may influence the alignment of the survey and administrative data, including age of the survey participant, program eligibility, and continuous program coverage.

4.1.7 Linked NCHS-CMS Medicare Match Probability Variable

Data linkages engender some uncertainty over which pairs represent true matches. For the survey data linked to CMS data, the probabilistic cut-off values used to determine which record pairs were considered a link (an inferred match) were set at a point that minimized both the type I error (false positives, or linked records that are not true matches) and the type II error (false negatives, or true matches that are not linked). For each candidate pair, the probability of match validity (PROBVALID) was computed. The PROBVALID cutoff is the threshold that produces the lowest total error (both type I and type II). However, because there are concerns that using pairs with low PROBVALID might be inappropriate for certain analyses of linked records, PROBVALID = 0.85 was established as the lowest threshold that will be used for the acceptance of links into datasets made available for external researchers.

Researchers can access PROBVALID to adjust linkage certainty by increasing the link acceptance cut-off scores or to conduct sensitivity analyses. For some analyses, it may be desirable to minimize type I error, which would be the result of using a value of PROBVALID closer to 1. Similarly, researchers may only want to include deterministic links, and could restrict the analysis to records with PROBVALID=1.

4.1.8 Considerations when Combining Data from Multiple NCHS-CMS Medicare Linkages

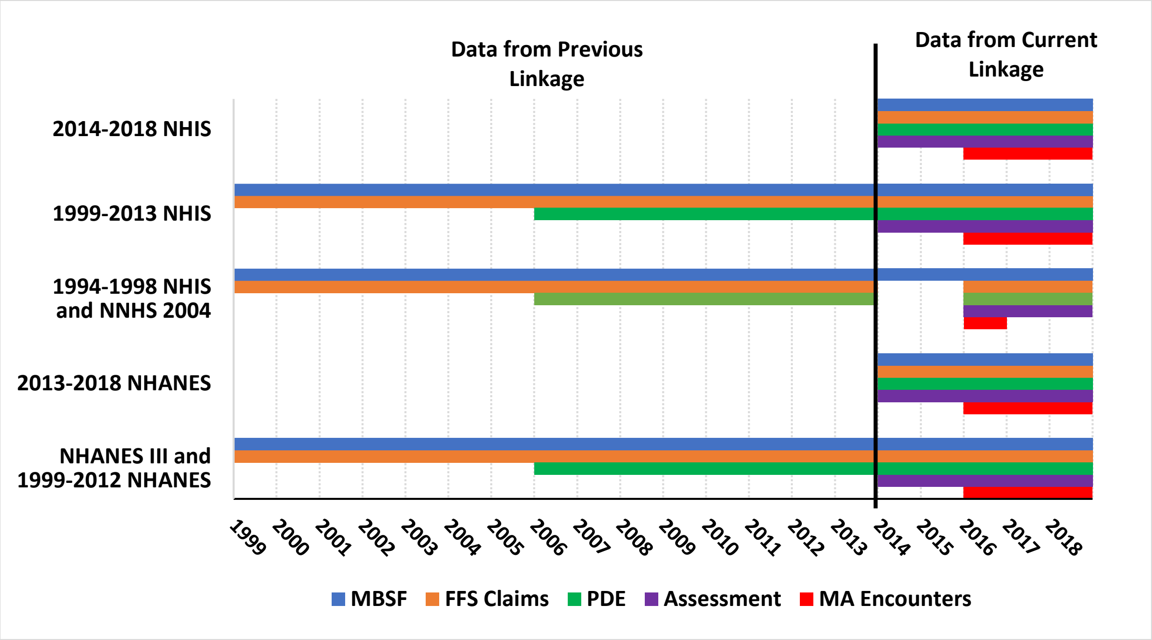
This report describes the linkage of NCHS survey data linked to 2014-2018 Medicare administrative data. Several NCHS surveys included in this most recent linkage have also been included in previous NCHS-CMS Medicare linkages[3]. Data from multiple Medicare linkages can be combined for approved research projects in the RDC. The linkage methodologies and linkage eligibility may differ between the time periods and should be taken into consideration when

[3] For certain NCHS surveys, linked Medicare enrollment and FFS claims data from 1991-1998 are also available. The format for these data will differ from the 1999-2013 and 2014-2018 linked CMS Medicare data files. Please contact the NCHS Data Linkage Team at (datalinkage@cdc.gov) for more information on data availability.

combining multiple years of linked data. For some surveys, such as NHANES III (1988-1994), there will be gaps of several years between the time the interview was conducted and the Medicare coverage period even when combining Medicare data from previous linkages.

As noted above, there may also be differences in the availability of Medicare data by survey depending on the survey participant's age, program eligibility and the year the survey was conducted. These variations in coverage periods should be taken into consideration by researchers when combining data across survey and Medicare coverage years.

Figure 1. Linked NCHS-CMS Medicare Data Availability for the linked NHIS and NHANES Surveys



NCHS recently linked the NCHS surveys listed in Figure 1 (above) with information from 2014-2018 Medicare enrollment, chronic condition flags, summarized annual health care utilization and cost data (i.e. MBSF), FFS health care claims and prescription drug event data, as well as MDS and OASIS assessment data.[4] Additionally, NCHS was able to conduct a first-time linkage with 2016–2018 MA Encounter data to enable analysis of Medicare beneficiaries enrolled in the MA program.

[4] In the initial April 2021 data release for surveys linked to the administrative data for calendar years 2014 and 2015, only MBSF data were available due to limited resources. Analyses of Medicare utilization in 2014–2015 were thus initially limited to summary measures of annual cost and utilization and indicators of the presence of selected chronic conditions available from the MBSF for 2014-2015. As of the updated April 2024 data release, 2014 and 2015 FFS health care claims, prescription drug event data, and MDS and OASIS data are now available for most linked surveys. Similarly, MA encounter data were initially available only for 2016, and the updated data release added MA Encounters for 2017 and 2018 for most linked surveys. No additional data was added for NHIS 1994-1998 and NNHS 2004.

In addition to the newly conducted linkage with 2014-2018 Medicare data, the 1994-2013 NHIS, the 2004 NNHS, and the 1999-2012 NHANES and NHANES III surveys were previously linked to 1999-2013 MBSF and FFS health care claims, creating the opportunity for extended follow-up analyses for survey participants enrolled in Medicare between 1999 and 2018. However, researchers should be aware that the linkage methodologies differ over time periods; in particular, the previous linkage to 1999–2013 MBSF and FFS health care claims used a deterministic match algorithm and did not include probabilistic matches.(6) Researchers may wish to request the estimated match probability variable PROBVALID (see Section 4.1.7) to estimate the sensitivity of derived estimates to the inclusion of probabilistic matches in the newly conducted linkage.

4.2 Analytic Considerations for Linked Medicare Data Files

Records for NCHS survey participants have been linked to the following CMS Medicare Data Files, which include enrollment data from the Master Beneficiary Summary File (MBSF), claims/encounter data from the FFS and MA files, and patient assessment data from long term care 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 1 are provided in Appendix I. The following sections address potential analytic considerations specific to each of the linked Medicare data files.

Table 1. List of CMS Medicare Data Files Linked to NCHS Survey Data

CMS Medicare Data Files	Years in Previous Linkage	Years in Current Linkage
Master Beneficiary Summary File (MBSF)	1999-2013	2014-2018
Medicare Fee-for Service (Claim Files)		
Inpatient (IP)	1999-2013	2014-2018*
Skilled Nursing Facility (SNF)	1999-2013	2014-2018*
Professional (Carrier)	1999-2013	2014-2018*
Outpatient (OP)	1999-2013	2014-2018*
Durable Medical Equipment (DME)	1999-2013	2014-2018*
Home Health Agency (HHA)	1999-2013	2014-2018*
Hospice	1999-2013	2014-2018*
Medicare Provider Analysis and Review File (MedPAR)	1999-2013	2014-2018*
Medicare Advantage (Encounter Files)		
Inpatient (IP)		2016-2018†
Skilled Nursing Facility (SNF)		2016-2018†
Professional (Carrier)		2016-2018†
Outpatient (OP)		2016-2018†

Durable Medical Equipment (DME)		2016-2018†
Home Health Agency (HHA)		2016-2018†
Medicare Part D Prescription Drug Event (PDE)	2006-2013	2014-2018*
Assessment Files		
Home Health Outcome and Assessment Information Set (OASIS)		2014-2018*
Long Term Care Minimum Data Set (MDS)		2014-2018*

* FFS Claims, PDE, and assessment files from the current linkage are available only for 2016–2018 for the following surveys: NHIS 1994–1998, NNHS 2004.

† MA Encounter files from the current linkage are available only for 2016 for the following surveys: NHIS 1994–1998, NNHS 2004.

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

The MBSF provides data on linked NCHS-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 NCHS survey data and 2016 MBSF will contain information for Medicare enrollment and summary health care utilization occurring in 2016.

Note: To properly construct linked NCHS-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 survey participant 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 survey participant 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.

Inclusion of Medicare Advantage (MA) Enrollees in Study Denominators

Until recently, CMS did not receive claims or encounter data for Medicare beneficiaries who enroll in MA. During the time covered by the linked data files, MA enrollment increased from approximately 30% of beneficiaries in 2014 to 34% in 2018 (7). A summary of the percentage of NCHS survey participants who were enrolled in a MA plan by year and survey can be found at https://www.cdc.gov/nchs/data/datalinkage/cms_medicare_linked_data_medicare_advantage_enrollment_tables.pdf.

At the time of the linkage only the 2016 Medicare Encounter Files were available. The updated data release in April 2024 added the 2017–2018 Medicare Encounter Files for most linked surveys. Researchers should consider the percentage of participants enrolled in MA when determining the feasibility and sample sizes of their proposed research projects using Medicare data for enrollment years without linked MA data (i.e. 2014-2015).

MA enrollment can be identified using the HMO indicators from the MBSF Base (A/B/C/D) segment. The file includes 12 HMO indicator variables (HMO_IND_01- HMO_IND_12), one for each month. During periods of MA 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 inpatient hospital services. Utilization of most Medicare-covered services is unobservable from Medicare claims data during periods of MA enrollment. Therefore, in general, studies based on analysis of claims data should exclude MA enrollees from their beneficiary samples, except for years where MA encounter data are available (2016–2018for the NCHS-Medicare linked files) and incorporated into the analysis.

For more information on how to create an analytic sample that excludes Medicare beneficiaries enrolled in a MA plan, refer to a document written by ResDAC https://www.resdac.org/articles/identifying-medicare-managed-care-beneficiaries-master-beneficiary-summary-or-denominator (accessed October 30, 2020).

Additional analytic considerations specific to analyzing data for MA enrollees are provided in Section 4.4.

Medicare Entitlement

The linked MBSF Base (A/B/C/D) segment also includes three variables indicating Medicare entitlement: original reason for entitlement, current reason for entitlement, and Medicare status code. A beneficiary's *original reason* for Medicare entitlement is found in the variable ENTLMT_RSN_ORIG. 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 October 30, 2020). As noted above (Section 4.1.4), the race and ethnicity information collected from the NCHS survey is self- or family respondent-reported and, thus, may be more accurate than the demographic information provided in the Medicare files. Therefore, when possible, the NCHS survey data should be used for demographic variables, such as race and ethnicity.

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 October 30, 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 October 30, 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 October 30, 2020).

4.2.1.4 MBSF File Year Indicator

The MBSF reference year can be found in the variables BENE_ENROLLMT_REF_YR and FILE_YEAR4. Please note that linked records for all years of Medicare enrollment are appended into a single file. It is possible that a single beneficiary can have MBSF records in multiple years. If this is the case, the beneficiary will appear multiple times 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 NCHS-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 October 30, 2020) or www.ccwdata.org (accessed October 30, 2020).

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

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

Durable medical equipment or DME can be billed through either a) the carriers who also process physician claims, or b) DME Regional Carriers (DMERCs), who process only DME claims. 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 NCHS survey participants 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 an FFS or MA plan.

4.3.4 Outpatient (OP) File

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

4.3.5 Inpatient (IP) File

Each record on this file represents a health care claim submitted for payment by inpatient hospital providers for reimbursement of facility costs incurred during the provision of inpatient care. Multiple claims records may be submitted for one hospital stay. Researchers interested in analyzing summarized information for inpatient stays rather than individual inpatient claims may wish to use the MedPAR file (described in Section 4.3.7) which summarizes individual inpatient claims at the stay level. Researchers interested in analyzing 2016–2018 inpatient data across the FFS and MA programs should use the FFS and MA Inpatient Files as there is currently no MedPAR type data file created to summarize Inpatient encounters at the stay level for the MA program.

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

4.3.6 Skilled Nursing Facility (SNF) File

Each claim record on this file represents a health care claim submitted for payment by a SNF for reimbursement of the provision of skilled nursing care. Multiple claims records may be submitted for one SNF stay. Medicare billing frequency guidance for SNFs requires SNFs to submit claims at least monthly. Researchers interested in analyzing claims information summarized at the stay level may wish to use the MedPAR file which summarizes individual SNF claims at the stay level (see Section 4.3.7). Researchers interested in analyzing 2016–2018 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 October 30, 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 NCHS Survey Linked Mortality Files (accessed October 30, 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 (MA) 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 there may be differences in the completeness of encounter data versus FFS claims data because of differences in the collection and payment purpose of MA encounter 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 capture 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).

CMS has published additional guidance to assist with analysis of Medicare encounter claims data which can be accessed at http://www.resdac.org (accessed October 30, 2020) or https://www2.ccwdata.org/documents/10280/19002246/ccw-medicare-encounter-data-user-guide.pdf (accessed October 30, 2020).

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 Affairs [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 October 30, 2020) or https://www.ccwdata.org (accessed October 30, 2020).

5 Additional Related Data Sources

Each of the NCHS surveys that have been linked to the Medicare data have also been linked to death information obtained from the NDI. The linked NDI mortality files provides the opportunity to conduct a vast array of outcome studies designed to investigate the association of a wide variety of health factors with mortality. More information about the NCHS-NDI linked mortality files can be found at https://www.cdc.gov/nchs/data-linkage/mortality.htm (accessed October 27, 2020).

NCHS has also previously linked to CMS Medicaid enrollment and claims data. Linkage of the NCHS survey participant data with the CMS Medicaid data provides the opportunity to study changes in health status, health care utilization and expenditures in low income families with children, the elderly and other vulnerable U.S. populations. For more information about the linked CMS Medicaid data, please see the data linkage website: https://www.cdc.gov/nchs/data-linkage/medicaid.htm (accessed October 20, 2020).

In addition, NCHS has linked to a separate set of data files containing information on patients diagnosed with ESRD obtained from the United States Renal Data System (USRDS) https://www.usrds.org/ (accessed October 27, 2020). The USRDS is a national data system, funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), designed to collect, analyze, and distribute information about ESRD in the United States. The linked ESRD data files can be used by researchers interested in conducting analysis specifically related to patients with ESRD. Although nearly all of the NCHS survey participants who were linked to USRDS records also linked to Medicare records, a small number of the linked USRDS records are not linked to Medicare records (less than 5%). For more information about the data available on the linked ESRD files, please refer to the documentation: https://www.cdc.gov/nchs/data-linkage/esrd.htm (accessed October 27, 2020).

NCHS survey data have also been linked to administrative data for the Department of Housing and Urban Development's (HUD) largest housing assistance programs: the Housing Choice Voucher program, public housing, and privately owned, subsidized multifamily housing. Combining the NCHS survey data with the linked Medicare and linked HUD data provides the opportunity to examine relationships between housing and health for the elderly population and persons with disability. For more information about the linked HUD data, please see the data linkage website: https://www.cdc.gov/nchs/data-linkage/hud.htm (accessed January 26, 2021).

Some of the NCHS surveys in the NCHS-CMS Medicare linkage have also been linked to administrative data from the Department of Veterans Affairs (VA). Researchers interested in outcomes related to Veterans may also request variables from the Linked NCHS-VA data files (accessed March 27, 2024). The Linked NCHS-VA data files include information on a wide range of health-related topics for Veterans, including Veteran status and utilization of VA benefit programs.

Data users may also request variables from the Linked CMS Medicaid, Linked ESRD, Linked HUD, Linked VA, or Linked NDI files (if mortality is the outcome of interest) in addition to the Linked NCHS–CMS Medicare Data Files. Each of these files can be merged with the Linked NCHS–CMS Medicare Data Files using the survey-specific unique participant identification variable (see Appendix III).

6 References

1. Miller, D.M., R. Gindi, and J.D. Parker, *Trends in record linkage refusal rates: Characteristics of National Health Interview Survey participants who refuse record linkage*. Presented at Joint Statistical Meetings 2011. Miami, FL., July 30–August 4.

2. Sayer, B. and C.S. Cox. *How Many Digits in a Handshake? National Death Index Matching with Less Than Nine Digits of the Social Security Number* in Proceedings of the American Statistical Association Joint Statistical Meetings. 2003.

3. Dahlhamer, J.M. and C.S. Cox, *Respondent Consent to Link Survey Data with Administrative Records: Results from a Split-Ballot Field Test with the 2007 National Health Interview Survey*. paper presented at the 2007 Federal Committee on Statistical Methodology Research Conference, Arlington, VA, 2007.

4. Fellegi, I.P., and A.B. Sunter, *A Theory for Record Linkage*. JASA, 1969. 40: 1183-1210.

5. Golden, C., et al., *Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare Medicaid Services*. Vital Health Stat 1, 2015(58): p. 1-53. https://www.cdc.gov/nchs/data/series/sr_01/sr01_058.pdf

6. National Center for Health Statistics, Division of Analysis and Epidemiology. The Linkage of National Center for Health Statistics Surveys to Medicare Enrollment and Claims Data (1999-2013) - Methodology and Analytic Considerations. December 2016. Hyattsville, Maryland.

7. CMS Program Statistics website https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CMSProgramStatistics/index (accessed December 12, 2020).

7 Additional Resources

NHANES–CMS Linked Data Tutorial
https://www.cdc.gov/nchs/tutorials/NHANES-CMS/index.htm (accessed November 6, 2020)

Information about the Medicare enrollment, claims/encounters, and assessment files was compiled from the following sources:

Centers for Medicare & Medicaid Services (CMS)
http://www.cms.gov/ (accessed November 6, 2020)

Chronic Conditions Data Warehouse
https://www2.ccwdata.org/web/guest/home/ (accessed November 6, 2020)

Research Data Assistance Center (ResDAC)
http://www.resdac.org/ (accessed November 6, 2020)

Appendix I: 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.[5] Additional information on each of the MBSF Segments may be found at https://www.cdc.gov/nchs/data-linkage/medicare-restricted.htm (accessed May 24, 2021).

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

[5] 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

claims processed as part of a CMS demonstration project.[6] Each segment is available as a separate file, but can be combined using the unique claim identification number (NCHS_CLM_ID), Medicare reference year (FILE_YEAR4) and unique survey participant identifier (see Appendix III).

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), Medicare reference year (FILE_YEAR4) and unique survey participant identifier (see Appendix III).

2.1 Inpatient (IP) Files

2.1.1 Fee-for-Service Inpatient File

The FFS IP File contains Medicare Part A final action claims from IP facilities. The FFS IP File contains data fields for ICD-10-CM/PCS codes, revenue center codes, dates of service, and payment information. Each record on this file contains the information from one health care claim. Episodes of care may encompass more than one health care claim. Additional information on the FFS IP File may also be found at https://www.cdc.gov/nchs/data-linkage/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

2.2 Skilled Nursing Facility (SNF) Files

2.2.1 Fee-for-Service SNF File

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

[6] 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)

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

Appendix I: Descriptions of Medicare Data Files

2.4 Outpatient (OP) Files

2.4.1 Fee-for-Service Outpatient File

The FFS OP File contains Medicare Part A final action claims from OP providers including: hospital OPDs, rural health clinics, renal dialysis facilities, OP rehabilitation facilities, comprehensive OP rehabilitation facilities, Federally Qualified Health Centers and community mental health centers. The FFS OP File contains data fields for ICD-10-CM/PCS codes, revenue center codes, dates of service, and payment information. Each record on this file contains the information from one health care claim. Episodes of care may encompass more than one health care claim. Additional information on the FFS OP File may also be found at https://www.cdc.gov/nchs/data-linkage/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

Appendix I: Descriptions of Medicare Data Files

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

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/medicare-restricted.htm (accessed May 24, 2021).

Appendix II: Detailed Description of Linkage Methodology

1 NCHS Survey and CMS Medicare Linkage Submission Files

Prior to the linkage of the NCHS survey and CMS Medicare 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 survey and Medicare 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 survey participant or enrollee):

- SSN validation[7]
- HICN[8]
- DOB
- Sex
- ZIP Code and State of residence
- First, 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 survey participants with multiple name parts, additional records were generated using each individual piece as a possible name value. Table 1 below provides two examples of how name information was used to generate alternate records, using

[7] Complete 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. For some surveys and survey years, only the last 4-digits (SSN4) were collected from survey participants.

[8] Complete 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). For some surveys and survey years, only the last 4-digits and a letter were collected from survey participants.

hypothetical data. For survey participant A, the first name was used to generate multiple records, and for survey participant B, the last name was used.

Appendix II: Detailed Description of Linkage Methodology

Table 1. Example of Alternate Record Generation using Name Fields

Survey Participant	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 survey participant PII was created for NCHS survey records and for CMS Medicare records. During this process, multiple submission file records were created for each survey participant/beneficiary to show all combinations of the recorded values for these fields. That is, if a survey participant 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 survey participant.

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). 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 survey participant/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 deterministic links were produced by pairing records with exactly the same SSN or HICN. 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 at least than 50% (for records that matched on the complete SSN or HICN) or 65% (for records that matched on SSN4 or the last 4-digits plus a letter of HICN), the linked pair was retained as a deterministic match (that is, a "true" match). In cases where this resulted in multiple matches for a single participant (i.e., in the case of alternate records), the record with the highest number of matching fields was retained. The collection of records resulting from the deterministic match was referred to as the 'truth deck.' Each of the deterministic pairs was assigned a probability of 1.

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)." (1) 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 NCHS survey and CMS EDB records as the training dataset. For more detailed information on this method please refer to *Learning Blocking Schemes for Record Linkage*.(2)

The machine learning algorithm generated 7 blocking passes to be used in the blocking scheme. Table 2 provides a specific breakdown of each blocking pass:

Table 2. Variables used for blocking and scoring to identify linked records

Block 1	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**Block 6**	**Block 7**
• Last Name • Full DOB • Zip Code	• Last Name • Full DOB • Sex	• First Name • Full DOB • Sex	• State • Full DOB • Sex	• Last Name • Month of Birth • State • Zip Code	• Last Name • First Name • State • Sex	• Year of Birth • Sex • State • Zip Code
Score 1	**Score 2**	**Score 3**	**Score 4**	**Score 5**	**Score 6**	**Score 7**
• First Name • Middle Initial • Sex	• First Name • Middle Initial • State • Zip Code	• Middle Initial • Last Name • State • Zip Code	• First Name • Middle Initial • Last Name • Zip Code	• First Name • Middle Initial • Year of Birth • Day of Birth • Sex	• Middle Initial • Month of Birth • Year of Birth • Day of Birth • Zip Code	• First Name • Middle Initial • Last Name • Month of Birth • Day of Birth

Appendix II: Detailed Description of Linkage Methodology

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 3.3 below), 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 2). 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.

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. 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 NCHS survey submission file and a simple random sample of 1% 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 NCHS survey 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 (3-5).

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:

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

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

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 survey participant’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.

$$N_{\widehat{matches},B} = \sum P_{EM,P,\widehat{B(M}atch)}$$

Continuing with the example, add the probabilities for every pair in the blocking pass:

$$N_{\widehat{matches,B}} = 0.87 + .0036 + \widehat{P_{EM,3,B}} + \ldots + \widehat{P_{EM,N_{Pairs,B},B}}$$

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.[9]

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.[10] This additional pair weight allows for more accurate adjudication of links where other PII may not provide a clear indication of match status.

When the last four digits of SSN[11] 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):

[9] 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 $\frac{1}{9{,}999} \approx 0.0001$

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

[11] 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.

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

For pairs that did not have an SSN on either the NCHS survey 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

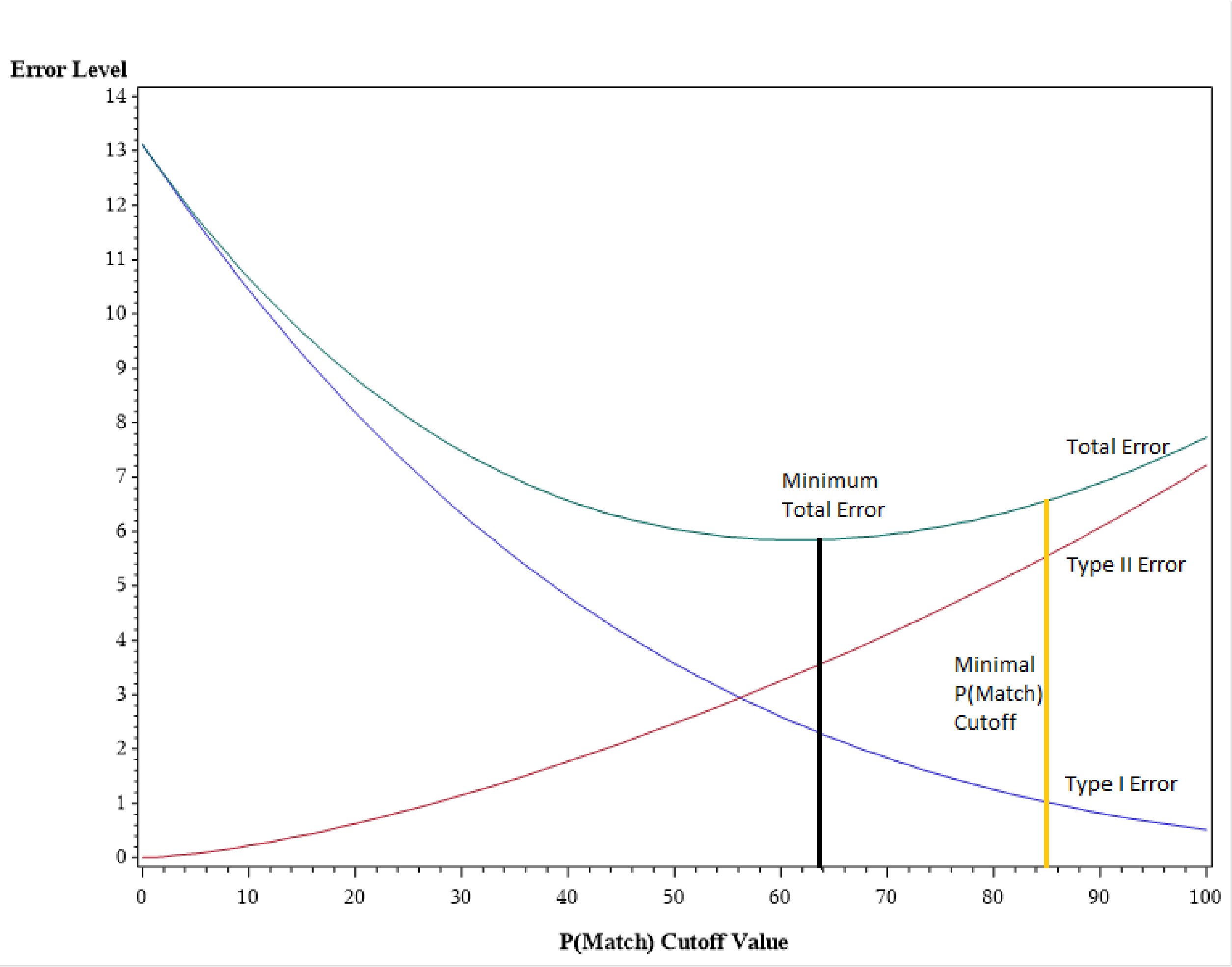
The estimated probabilities were used to measure Type I error. For the probabilistic records, the estimated match probabilities represented the probability that the NCHS survey record was a match to the CMS EDB record. In other words, if a link had an estimated probability of 0.98, then it was understood 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 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.

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 fewer 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 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) = 0.85 was established as the lowest threshold that will be used for the acceptance of links into datasets made available for external researchers. The vast majority (>95%) of records with probabilistic links have P(Match) >= 95%.

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 each NCHS survey participant (if more than one link 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 NCHS survey-CMS Medicare Data linkage were 0.04% and 0.33%, respectively. These differed slightly for those with SSN9 and SSN4s (data not shown).

Appendix II: Detailed Description of Linkage Methodology

5 References

1. Christen, P. *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).

2. Michelson, Matthew, and Craig A. Knoblock. *Learning Blocking Schemes for Record Linkage*. 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).

3. 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.

4. 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.

5. Resnick, D.M., Mirel, L.B., Roemer, M.I., Campbell, S.R. 2020. Adjusting Match Weights to Partial Levels of String Agreement in Data Linkage. In JSM Proceedings, Government Statistics Section. Alexandria, VA: American Statistical Association.

Appendix III: Merging Linked NCHS-CMS Medicare Files with NCHS Survey Data

The data provided on the 1994-2018 NHIS, 1999-2018 NHANES, NHANES III, and the 2004 NNHS linked CMS Medicare files can be merged with the NCHS restricted and public use survey data files using the unique survey specific Public Identification number (PUBLICID/SEQN/RESNUM).

Note: At this time the linked CMS Medicare data files are only available for research use through the NCHS restricted access data center (RDC). Approved RDC researchers may choose to provide their own analytic files created from public use survey files to the RDC. Therefore, it is important for researchers to include survey specific Public Identification number on any analytic files sent to the RDC. The RDC will merge data (using PUBLICID, SEQN or RESNUM) from the linked CMS Medicare files to the analyst's file. The merged file will be held at the RDC and made available for analysis.

Information on how to identify and/or construct the NCHS survey specific PUBLICID, SEQN or RESNUM is provided below.

1 National Health Interview Survey (NHIS), 1994-2018

1.1 **NHIS, 1994**

Variable	Public-use Location	Length	Description
YEAR	3-4	2	Year of interview
QUARTER	5	1	Calendar quarter of interview
PSUNUMR	6-8	3	Random recode of PSU
WEEKCEN	9-10	2	Week of interview within quarter
SEGNUM	11-12	2	Segment number
HHNUM	13-14	2	Household number within quarter
PNUM	15-16	2	Person number within household

Note: Concatenate all variables to get the unique person identifier.

SAS example:

```
length publicid $14;
PUBLICID = trim(left(YEAR||QUARTER||PSUNUMR||WEEKCEN||SEGNUM||HHNUM||PNUM));
```

Stata example: (note this will convert the variables to string variables)

```
egen PUBLICID = concat(YEAR QUARTER PSUNUMR WEEKCEN SEGNUM HHNUM PNUM)
```

Appendix III: Merging Linked NCHS-CMS Medicare Files with NCHS Survey Data

1.2 **NHIS, 1995-1996**

Variable	Public-use Location	Length	Description
YEAR	3-4	2	Year of interview
HHID	5-14	10	Household ID number
PNUM	15-16	2	Person number within household

Note: Concatenate all variables to get the unique person identifier.

SAS example:

```
length publicid $14;
PUBLICID = trim(left(YEAR||HHID||PNUM));
```

Stata example: (note this will convert the variables to string variables)

```
egen PUBLICID = concat(YEAR HHID PNUM)
```

1.3 **NHIS, 1997-2003**

Variable	Public-use Location	Length	Description
SRVY_YR	3-6	4	Year of interview
HHX	7-12	6	Household number
FMX	13-14	2	Family number
PX	15-16	2	Person number within household

Note: Concatenate all variables to get the unique person identifier.

SAS example:

```
length publicid $14;
PUBLICID = trim(left(SRVY_YR||HHX|| FMX||PX));
```

Stata example: (note this will convert the variables to string variables)

```
egen PUBLICID = concat(SRVY_YR HHX FMX PX)
```

*The person identifier was called PX in the 1997-2003 NHIS and FPX in the 2004 (and later) NHIS; users may find it necessary to create an FPX variable in the 2003 and earlier datasets (or PX in later datasets).

Appendix III: Merging Linked NCHS-CMS Medicare Files with NCHS Survey Data

1.4 **NHIS, 2004**

Variable	Public-use Location	Length	Description
SRVY_YR	3-6	4	Year of interview
HHX	7-12	6	Household number
FMX	13-14	2	Family number
FPX	15-16	2	Person number within household

Note: Concatenate all variables to get the unique person identifier.

SAS example:

```
length publicid $14;
PUBLICID = trim(left(SRVY_YR||HHX||FMX||FPX));
```

Stata example: (note this will convert the variables to string variables)

```
egen PUBLICID = concat(SRVY_YR HHX FMX FPX)
```

1.5 **NHIS, 2005-2018**

Variable	Public-use Location	Length	Description
SRVY_YR	3-6	4	Year of interview
HHX	7-12	6	Household number
FMX	16-17	2	Family number
FPX	18-19	2	Person number within household

Note: Concatenate all variables to get the unique person identifier.

SAS example:

```
length publicid $14;
PUBLICID = trim(left(SRVY_YR||HHX||FMX||FPX));
```

Stata example: (note this will convert the variables to string variables)

```
egen PUBLICID = concat(SRVY_YR HHX FMX FPX)
```

Appendix III: Merging Linked NCHS-CMS Medicare Files with NCHS Survey Data

2 National Health and Nutrition Examination Survey (NHANES), 1999-2018

Item	Length	Description
SEQN	6	Participant identification number

All of the NHANES public-use data files are linked with the common survey participant identification number (SEQN). Merging information from multiple NHANES Files to the NHANES-CMS Medicare linked files using this variable ensures that the appropriate information for each survey participant is linked correctly.

3 Third National Health and Nutrition Examination Survey (NHANES III)

Item	Length	Description
SEQN	5	Participant identification number

All of the NHANES III public-use data files are linked with the common survey participant identification number (SEQN). Merging information from multiple NHANES III Files to the NHANES III-CMS Medicare linked files using this variable ensures that the appropriate information for each survey participant is linked correctly.

Appendix III: Merging Linked NCHS-CMS Medicare Files with NCHS Survey Data

4 National Nursing Home Survey (NNHS), 2004

Item	Length	Description
RESNUM	6	Resident Record (Case) Number

All of the 2004 NNHS public-use data files are linked with the common resident record (case) number (RESNUM). Merging information from the 2004 NNHS Files to the 2004 NNHS-CMS Medicare linked files using this variable ensures that the appropriate information for each survey participant is linked correctly.