The Linkage of National Center for Health Statistics Survey Data to 2019–2021 Medicare Enrollment and Claims/Encounters Data: Linkage Methodology and Analytic Considerations

Data Release Date: December 10, 2025

Document Version Date: September 23, 2025

Division of Analysis and Epidemiology

National Center for Health Statistics

Centers for Disease Control and Prevention

datalinkage@cdc.gov

Suggested citation:

National Center for Health Statistics, Division of Analysis and Epidemiology. The Linkage of National Center for Health Statistics Survey Data to 2019–2021 Medicare Enrollment and Claims/Encounters Data: Linkage Methodology and Analytic Considerations. December 2025. Hyattsville, Maryland.

Available from the following address:

https://www.cdc.gov/nchs/linked-data/medicare/index.html

Contents

List of Acronyms

AMA, American Medical Association

CCW, Chronic Conditions Warehouse

CMS, Center for Medicare & Medicaid Services

CPT-4, Current Procedural Terminology, 4th Edition

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

MBI, Medicare Beneficiary Identifier

MBSF, Master Beneficiary Summary File

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

OP, outpatient

OTC, over-the-counter

PDE, prescription drug event

PDP, prescription drug plan

PII, personally identifiable information

RDC, Research Data Center

ResDAC, Research Data Assistance Center

SAF, standard analytic file

SNF, skilled nursing facility

SSN, Social Security number

VA, Veteran Affairs

VRDC, Virtual Research Data Center

1 Introduction

As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to provide statistical information that can be used to guide actions and policy to improve the health of the American people. In addition to collecting and disseminating the Nation's official vital statistics, NCHS conducts several population-based surveys and healthcare establishment surveys that provide information on a wide-range of health-related topics, that often lack information on longitudinal outcomes.

Through its Data Linkage Program, NCHS has been able to expand the analytic utility of 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. The linked NCHS-Medicare files combine health and socio-demographic information from the surveys with enrollment and claims/encounters 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.

Under an interagency agreement between NCHS and the Centers for Medicare & Medicaid Services (CMS) several NCHS surveys have been linked to 2019–2021 Medicare enrollment, Medicare Part A and B beneficiary fee-for-service (FFS) health care claims, Medicare Advantage (MA) beneficiary encounter data, and Medicare Part D prescription drug events.

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

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, Claims/Encounters and Assessment Data (2014–2018): Linkage Methodology and Analytic Considerations* https://www.cdc.gov/nchs/data/datalinkage/NCHS-Medicare-Linkage-Methodology-and-Analytic-Considerations.pdf

- *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

- *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

2 Data Sources

2.1 National Center for Health Statistics, Survey Data

NCHS has recently linked the following surveys to 2019–2021 Medicare enrollment, FFS claims, MA encounter data, and prescription drug events:

- 1999–2021 National Health Interview Survey (NHIS)
- 1999–2018 and 2017–March 2020 Pre-Pandemic Continuous National Health and Nutrition Examination Survey (NHANES)
- Third National Health and Nutrition Examination Survey (NHANES III)

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. 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 in section 3.1. NHIS implemented its most recent content and structure redesign in 2019. For detailed information on the NHIS's contents and methods, refer to the NHIS website, https://www.cdc.gov/nchs/nhis/index.html

NHANES is a nationally representative cross-sectional survey designed to monitor the health and nutritional status of the civilian noninstitutionalized U.S. population. The NHANES sample is selected through a complex, multistage probability design. The sample design includes oversampling to obtain reliable estimates of health and nutritional estimates for population subgroups. The survey consists of interviews conducted in participants' homes and standardized physical examinations conducted in mobile examination centers.

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 two 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/.

2.2 Centers for Medicare & Medicaid Services, Medicare Data

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

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

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

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

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

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

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

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

3 Linkage of NCHS Survey Data with 2019–2021 Medicare Records

3.1 Linkage Eligibility

The linkage of NCHS survey participants to their Medicare enrollment and claims/encounters 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)[1] 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 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 the 2019–2021 Medicare linkage, only adult survey participants are included on the linked data files available to researchers. Since the majority of Medicare beneficiaries are age 65 and older, there were only a small number of child survey participants with linked Medicare records. Due to the limited analytic utility of these records and the increased disclosure risk they present, all child survey participants were classified as ineligible for linkage and their records were excluded from the linked Medicare files.

For many of the surveys initiated prior to and during 2007, including 1999–2006 NHIS, 1999–2008 NHANES, and NHANES III, a refusal by the survey participant to provide an SSN or Medicare Health Insurance Claim Number (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.[2] In attempt to address declining linkage eligibility rates, NCHS introduced new procedures

[1] 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.

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.[3] The research assessed algorithms using the last four and last six digits of SSN. The results provided support for changes in how NHIS collected SSN and HICN for linkage.[4]

Beginning in 2007, NHIS started requesting only the last four digits of SSN and HICN (plus the 1-3 character suffix) 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.

Beginning in 2018, CMS transitioned from SSN-based HICNs to randomly-generated alphanumeric Medicare Beneficiary Identifiers (MBIs) for administering Medicare transactions. As a result, in 2019, NHIS asked survey respondents enrolled in Medicare to provide either the last four digits of HICN (plus the 1-3 character suffix) or the last four characters of MBI. In 2020, NHIS began asking survey participants enrolled in Medicare to provide solely the last four characters of MBI. As with previous survey years (2007 forward), NHIS participants who refused to provide an MBI were asked for their explicit permission to link to administrative records without SSN or Medicare number.

For continuous NHANES, the informed consent procedures changed 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 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 or complete HICN were given the option to provide only the last four characters of either identification number. In 2019–March 2020, survey participants could provide either HICN or MBI as the Medicare identifier.

For NCHS adult survey participants (age 18 and older at the time of survey for NHIS and age 20 and older at the time of the survey for NHANES) who provided consent for linkage (as described above), linkage was attempted for those who had at least two of the following three identifiers present:

- valid SSN[2]
- valid date of birth (month, day, and year)[3]
- valid name (first, middle initial, and last)[4]

For example, if the PII on the survey record had no SSN, a full name, and only the year of birth, the record would be considered ineligible for linkage, as only one of the criteria (i.e., that for name) was met.

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

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

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

3.2 Overview of Linkage Methodology

This section outlines steps that were used to link the NCHS survey data with 2019–2021 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 linkage-eligible 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), 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.[5] Following this, a selection process was implemented with the goal of selecting pairs that represented the same individual between the two data sources. The following steps were implemented:

1. Deterministic linkage joined records on exact SSN (9 digits or last 4 digits) and validated links by comparing other identifying fields (i.e., first name, last name, day of birth, etc.)
2. Probabilistic linkage identified likely matches, or links, between all records. All records were probabilistically linked and scored as follows:
    a. Formed pairs via blocking
    b. Scored pairs
    c. Modeled probability – assigned estimated probability that pairs are matches
3. Pairs were selected which were believed to represent the same individual between data sources (i.e. they are a match)
    a. Deterministic matches (from step 2) were assigned a match probability of 1
    b. Record pairs selected from the probabilistic match (step 3) were assigned the model match probability. Record pairs with a match probability above the probability cut-off value were determined to be matches.

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

3.3 Linkage Rates

The linkage rates for the 2019–2021 NCHS-CMS Medicare linkage are provided in tables which can be accessed from this website: https://www.cdc.gov/nchs/linked-data/medicare/methods.html. For each linked NCHS survey, the tables present the total survey sample size for adults, 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 adult 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 current Medicare linkage period.

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 (NHIS) or 20–64 years (NHANES), and 65 years and older. Age was defined as the survey participant's age at interview. For the earliest linked surveys, the match rate may be lower among adults aged 65 and older at interview compared with adults under age 65 at interview. This is because some adults under age 65 at interview have aged into Medicare eligibility before the end of the administrative data years, while some adults aged 65 and older 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/ 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/.

4.1.2 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 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 Linked NCHS-CMS Medicare Data files. The unique survey identifiers are survey-specific and may be constructed differently across survey years. Please refer to Appendix III for guidance on identifying and constructing (if necessary) the appropriate identification variable for merging survey data and the Linked NCHS-CMS Medicare Data files.

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

4.1.3 Survey Variables to Include 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 (if any) and for use with the variables from the linked CMS data files. The restricted variables from NCHS surveys and the exact variables from the linked CMS 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 NCHS linked mortality files (LMFs) 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 LMFs can be found at https://www.cdc.gov/nchs/linked-data/mortality-files/index.html.
- 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 include 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. Linkage-eligibility adjusted weights are provided. However, researchers who wish to apply a different adjustment method should include the appropriate original sample weight(s) from the public-use survey files.
- Demographic information about survey participants from the NCHS survey— For variables such as race and ethnicity (Hispanic origin), 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: For more information about **variables from the linked CMS Medicare files** that should be considered for inclusion in all RDC proposals, see Section 4.2.1.1 MBSF Base Segment File (Medicare Parts A/B/C/D). 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.

4.1.4 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. NCHS has included eligibility-adjusted weights in the **Match Status and Weights file** using this approach. 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*[6] and in Assessing Linkage Eligibility Bias in the National Health Interview Survey.[7]

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. If researchers wish to conduct their own weight adjustment, this can be done by requesting the appropriate sample weight(s) from the public-use survey files and using the MEDICARE_MATCH_1921 variable from the Match Status and Weights file to determine linkage eligibility.

The choice of which adjusted sample weight to use depends on the analysis and, more specifically, on the variables used in the analyses and the survey years included. Below are important considerations for the two surveys.

For **NHIS**: Since all persons in the household sampled in the 1999-2006 NHIS were potentially eligible for linkage, eligibility-adjusted analyses of 1999-2006 NHIS should incorporate the person weights (FA_WGT_ADJ), or the sample adult weights (SA_FA_WGT_ADJ) (if analytic variables are based on sample adult file). As only sample adults were potentially eligible for linkage in the 2007-2021 NHIS, eligibility-adjusted analyses of 2007-2021 NHIS sample adult participants should incorporate the adjusted sample adult weights. For NHIS 2020, researchers are advised to review the NHIS 2020 documentation for discussion of changes to field procedures in response to the COVID-19 pandemic, the three analytic data files available for sample adults, and the appropriate weight to use.

For **Continuous NHANES**: Analyses should incorporate either the eligibility-adjusted interview weights (ADJ_INTWT) or, if analytic variables are based on data obtained during the MEC examination, the adjusted MEC examination weights (ADJ_MECWT). For analyses of the combined 1999–2000 and 2001–2002 survey years, adjusted 4-year interview weights (ADJ_4YR_INTWT) and examination weights (ADJ_4YR_MECWT) are also available. Researchers are advised to consult the NHANES analytic guidelines for more information about constructing weights when analyzing multiple survey cycles and selecting the appropriate sample weight for analysis. Eligibility-adjusted weights have not been included for other NHANES subsamples (e.g. fasting subsample or dietary subsamples); researchers may wish to conduct their own weight adjustments for analyses using these weights.

For **NHANES III**: Eligibility-adjusted weights have not been included for NHANES III participants due to the high linkage eligibility rate for this survey.

4.1.5 Linked NCHS-CMS Medicare Match Probability Variable

The **Match Status and Weights file** also contains information on the estimated probability of match validity (PROBVALID). An estimated probability of match validity was computed for each candidate pair and compared against a probability cut-off value to determine which pairs were links (an inferred match). For additional discussion on how PROBVALID was estimated, see Appendix II Sections 3.3 and 3.4. NCHS used a probability cut-off value which minimized the total estimated counts of Type I error (false positive links) and Type II error (false negative links).

In the NCHS - 2019–2021 CMS Medicare linkages, NCHS used a probability cut-off value of 0.85 to determine final match status. Candidate pairs with a PROBVALID that exceeded the probability cut-off (i.e., PROBVALID>0.85) were considered linked. For additional discussion on probability cut-off value determination and record selection, please see Appendix II Section 4. For some analyses, it may be desirable to reduce the Type I error. To do this, researchers should increase the probability cut-off value to a value closer to 1.0. Researchers who only want to include deterministic links could restrict the analysis to record with PROBVALID=1. Researchers wishing to increase the probability cut-off value should request PROBVALID in their RDC proposal. Note, the probability cut-off value cannot be decreased from 0.85 as pairs estimated with lower match probability are not made available to researchers.

4.1.6 Temporal Alignment of Survey and Administrative Data

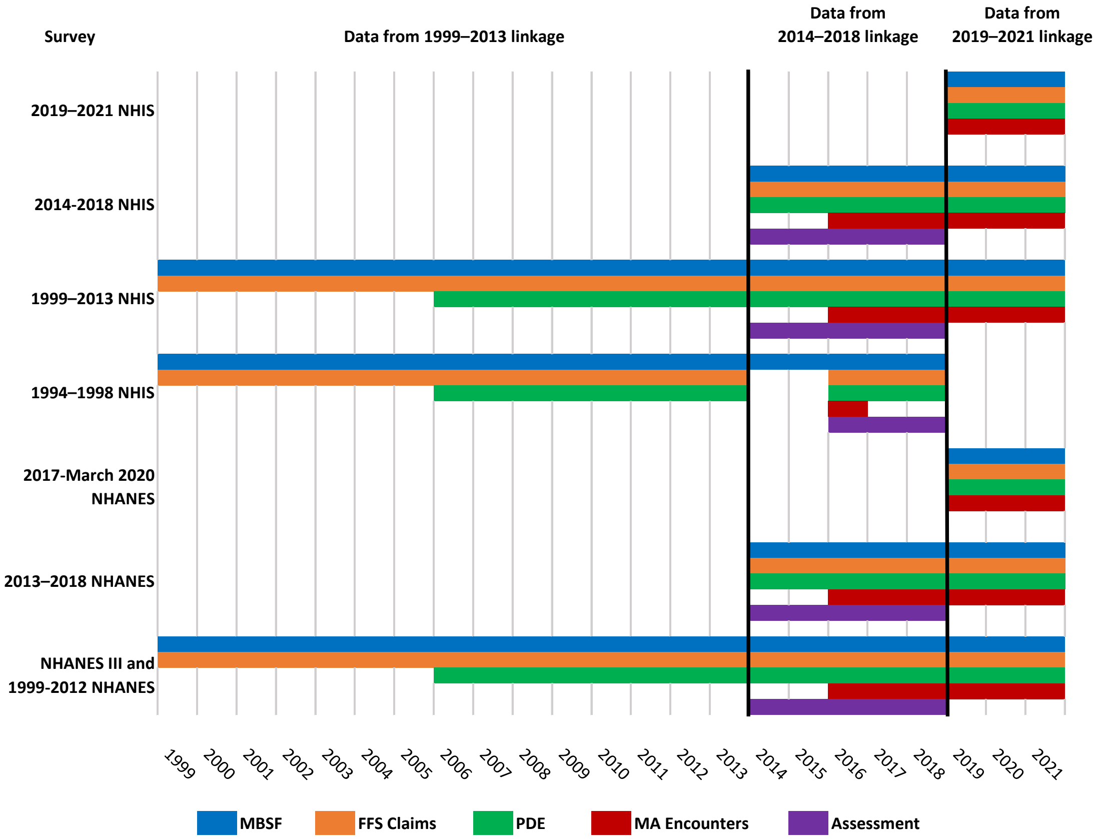
NCHS surveys have been linked to multiple years of Medicare administrative data. Depending on the survey year, Medicare data may be available for survey participants at the time of the survey, as well as before or after the survey period. Several factors may influence the alignment of the survey and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

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

This report describes the linkage of NCHS survey data linked to 2019–2021 Medicare administrative data. Several NCHS surveys included in this most recent linkage have also been included in previous NCHS-CMS Medicare linkages.[5] 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 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.

[5] 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, 2014–2018 and 2019–2021 linked CMS Medicare data files. Please contact the NCHS Data Linkage Team at (datalinkage@cdc.gov) for more information on data availability.

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



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.

4.2 Analytic Considerations for Linked Medicare Data Files

Records for NCHS survey participants have been linked to 2019–2021 records from the following CMS Medicare Data Files:

- Master Beneficiary Summary File (MBSF)
  - Base (Medicare Parts A/B/C/D) Segment
  - Cost and Use Segment
  - Chronic Conditions Segments
- Part D Prescription Drug Event (PDE)
- Medicare Provider Analysis and Review File (MedPAR)
- Fee-for-Service (Claim files)
  - Inpatient (IP)
  - Skilled Nursing Facility (SNF)

    - Professional (Carrier)
    - Outpatient (OP)
    - Durable Medical Equipment (DME)
    - Home Health Agency (HHA)
    - Hospice
- Medicare Advantage (Encounter Files)
    - Inpatient (IP)
    - Skilled Nursing Facility (SNF)
    - Professional (Carrier)
    - Outpatient (OP)
    - Durable Medical Equipment (DME)
    - Home Health Agency (HHA)

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

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

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 2019 MBSF will contain information for Medicare enrollment and summary health care utilization occurring in 2019.

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

Creating Medicare Study Denominators

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.

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 Medicare Advantage Organization (MAO).

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

For more information on how to create an analytic sample that excludes Medicare beneficiaries enrolled in a MA plan, refer to a document written by ResDAC https://www.resdac.org/articles/identifying-medicare-managed-care-beneficiaries-master-beneficiary-summary-or-denominator. Additional analytic considerations specific to analyzing data for MA enrollees are provided in Section 4.2.3.

Medicare Entitlement

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

- A beneficiary's *original reason* for Medicare entitlement is found in the variable ENTLMT_RSN_ORIG. Knowing a beneficiary's original reason for entitlement can be useful for identifying which aged beneficiaries were formerly entitled (i.e., prior to age 65) to Medicare due to a qualifying disability., Possible values include: Old Age and Survivors Insurance (OASI), Disability Insurance Benefits (DIB) and ESRD.
- A beneficiary's *current reason* for Medicare entitlement is found in the variable ENTLMT_RSN_CURR. Possible values include: OASI, DIB and ESRD.
- The variables MDCR_STATUS_CODE_01 - MDCR_STATUS_CODE_12 specify the monthly status of the beneficiary's entitlement to Medicare benefits. Possible values include: Aged without ESRD, Aged with ESRD, Disabled without ESRD, Disabled with ESRD, and ESRD only.

Race and Ethnicity

The linked MBSF Base (A/B/C/D) Segment provides two race and ethnicity variables BENE_RACE_CD, which is the variable reported in the CMS administrative claims data system, and RTI_RACE_CD, which contains race and ethnicity codes imputed through the use of an algorithm developed by the Research Triangle Institute (RTI) to improve the accuracy of race and ethnicity data included in the administrative claims data system. More detailed information regarding the RTI algorithm can be found at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4195038.

As noted above (in Section 4.1.3), the race and Hispanic origin information collected from the NCHS survey, which is self- or family respondent-reported, also is available for the linked files.

4.2.1.2 MBSF Cost and Use Segment

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

4.2.1.3 MBSF Chronic Conditions Segments

The CMS Medicare MBSF Chronic Conditions segments include variables indicating whether each Medicare FFS-enrolled beneficiary has claims indicating the presence of multiple specific chronic conditions. The 2019–2021 linked survey data includes two versions of the Chronic Conditions categories: the 27 CCW Chronic Conditions file and the 30 CCW Chronic Conditions file, which uses enhanced algorithms. CMS provides additional information about the methodology used to assign chronic condition flags to Medicare beneficiaries on their website (https://www2.ccwdata.org/web/guest/condition-categories-chronic) and in the Chronic Conditions File Enhancement White Paper (available at https://www2.ccwdata.org/documents/10280/19002256/ccw-condition-categories-impact-of-transition-from-27-to-30.pdf).

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

4.2.1.4 MBSF Other Chronic or Potentially Disabling Conditions

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

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

4.2.1.5 MBSF File Year Indicator

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

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

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

The data provided on the linked 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 a Medi-gap policy. Therefore, the total amount spent for a given health care service may not be captured by relying on the Medicare FFS claims payment data alone. CMS has published additional guidance to assist with analysis of Medicare FFS claims data in the CCW Medicare Data User Guide.

4.2.2.1 Carrier File

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

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

The FFS Carrier File contains DME claims processed by payment contractors who also process physician claims. The DME line items included on the FFS Carrier File can be identified by Claim Type Code (NCH_CLM_TYPE_CD) equal to 72. DME claims processed through DME regional carriers are found on the FFS DME Files, not on the Carrier File. DME claims on the Carrier File are for separate services. For additional information on DME regional carrier claims, see the DME File description in Section 4.2.2.2.

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

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

4.2.2.2 Durable Medical Equipment (DME) File

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

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

4.2.2.3 Hospice File

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

4.2.2.4 Outpatient (OP) File

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

4.2.2.5 Inpatient (IP) File

Each record on this file represents a health care claim submitted for payment by inpatient hospital providers for reimbursement of facility costs incurred during the provision of inpatient care. Multiple claims records may be submitted for one hospital stay. Researchers interested in analyzing summarized information for inpatient stays rather than individual inpatient claims may wish to use the MedPAR file (described in Section 4.2.2.7) which summarizes individual inpatient claims at the stay level. Researchers interested in analyzing inpatient data across the FFS and MA programs should use the FFS and MA Inpatient Files as there is currently no MedPAR type data file created to summarize Inpatient encounters at the stay level for the MA program.

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

4.2.2.6 Skilled Nursing Facility (SNF) File

Each claim record on this file represents a health care claim submitted for payment by a SNF for reimbursement of the provision of skilled nursing care. Multiple claims records may be submitted for one SNF stay. Medicare billing frequency guidance for SNFs requires SNFs to submit claims at least monthly. Researchers interested in analyzing claims information summarized at the stay level may wish to use the MedPAR file which summarizes individual SNF claims at the stay level (see Section 4.2.2.7). Researchers interested in analyzing SNF data across the FFS and MA programs should use the FFS and MA SNF Files as there is currently no MedPAR type data file created to summarize SNF encounters at the stay level for the MA program.

4.2.2.7 Medicare Provider Analysis and Review (MedPAR) File

The MedPAR file creates a single summarized record for each hospital or SNF stay, containing information on ICD-10-CM/PCS codes, admission, discharge, and procedure dates from the individual IP and SNF final action claims. Information regarding charges for IP or SNF services are more highly aggregated in MedPAR than those provided in the Inpatient and SNF Claims Files. Each MedPAR record may represent one IP or SNF claim or multiple claims, depending on the length of a beneficiary's stay and the amount of services billed throughout the stay. Researchers interested in the more granular detail of individual IP or SNF claims should use the FFS IP or SNF Claims Files for their analyses.

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

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

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

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

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

There are 2 types of encounter data records: Encounter Data Records and Chart Review Records.

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

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

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

4.2.4 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.

CMS has published additional guidance to assist with analysis of Medicare prescription drug data in the CCW

Medicare Part D Data User Guide.

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 mortality files provide the opportunity to conduct a vast array of outcome studies designed to investigate the association of a wide variety of health factors with mortality. For more information about the NCHS linked mortality files, please see the data linkage website: https://www.cdc.gov/nchs/linked-data/mortality-files/index.html.

NCHS has 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/linked-data/medicaid/index.html.

NCHS survey data have 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/linked-data/hud/index.html.

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. 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. For more information about the linked VA data, please see the data linkage website: https://www.cdc.gov/nchs/linked-data/va/index.html.

Data users may also request variables from the Linked CMS Medicaid, Linked HUD, Linked VA, or Linked Mortality files 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] CMS Medicare Enrollment Dashboard. https://data.cms.gov/tools/medicare-enrollment-dashboard

[2] 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.

[3] 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.

[4] 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.

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

[6] 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

[7] Aram J, Zhang C, Golden C, Zelaya CE, Cox CS, Ye Y, Mirel LB. Assessing linkage eligibility bias in the National Health Interview Survey. National Center for Health Statistics. Vital Health Stat 2(186). 2021. DOI: https://dx.doi.org/10.15620/cdc:100468.

7 Additional Resources

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/

- Chronic Conditions Data Warehouse

  https://www2.ccwdata.org/web/guest/home/

- Research Data Assistance Center (ResDAC)

  http://www.resdac.org/

Appendix I Descriptions of Medicare Data Files

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

1 Master Beneficiary Summary File (MBSF)

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

2 Standard Analytic Files (SAFs)

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

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

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

The Carrier and DME files share similar formats. Each file consists of 1) a base claims segment, containing demographic information and diagnosis codes as well as billing and payment amounts associated with a non-institutionalized claim; 2) a line items segment that includes the specific billing and payment amounts for each line item included within the base claim; and 3) a demonstration code segment. The base claim, line item, and demonstration code segments are available as separate files but can be combined using the unique claim identification number (NCHS_CLM_ID) or encounter join key (NCHS_ENC_JOIN_KEY), Medicare reference year (FILE_YEAR4) and unique 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.

2.1.2 Encounter Inpatient File

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

2.2 Skilled Nursing Facility (SNF) Files

2.2.1 Fee-for-Service SNF File

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

2.2.2 Encounter SNF File

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

2.3 Carrier Files

2.3.1 Fee-for-Service Carrier File

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

Section 4.3.2) 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.

2.3.2 Encounter Carrier File

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

2.4 Outpatient (OP) Files

2.4.1 Fee-for-Service Outpatient File

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

2.4.2 Encounter Outpatient File

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

2.5 Durable Medicare Equipment (DME) Files

2.5.1 Fee-for-Service DME File

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

2.5.2 Encounter DME File

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

2.6 Home Health Agency (HHA) Files

2.6.1 Fee-for-Service HHA File

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

2.6.2 Encounter HHA File

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

2.7 Hospice File

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

3 Medicare Provider Analysis and Review (MedPAR) File

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

4 Medicare Part D Prescription Drug Event (PDE) File

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

Appendix II Detailed Description of Linkage Methodology

1 NCHS Survey and CMS Medicare Linkage Submission Files

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

- SSN (validated)[6]
- DOB (month, day, and year)
- Sex
- Zip Code and State of residence
- First, middle initial, and last name[7]

The collection of HICN and MBI changed over time and varied across survey and survey years. Due to the inconsistencies in the completeness and quality of these identifiers, they were not used in the linkage process. However, 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 did not have a valid SSN or if the extracted SSN from HICN differed from the SSN provided during data collection, the extracted SSN value was retained as an additional SSN value to be used in the linkage process.

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

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

To increase the likelihood of finding a link, multiple or alternate submission records could be generated for each linkage eligible record in the NCHS survey participant and CMS Medicare submission files based

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

[7] Some survey records include maiden name or father's surname which, when different from the recorded last name, were treated as an alternate last name that was used to create an alternate survey submission record.

on variation of the linkage variables. Similar to the cleaning process, a more elaborate routine was used to generate alternate records involving the name fields. Alternate records were generated according to the following rules.

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

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

Table I. Example of Alternate Record Generation using Name Fields

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

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

Submission files, which combined the cleaned and validated PII fields, were created separately for NCHS survey records and for CMS Medicare records. During this process, multiple submission records were created for each survey participant/Medicare 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 dates-of-birth recorded and each of the remaining fields had only one variant, then a total of six submission records would have been created for the survey participant (see Table II for example). Submission records that did not meet the eligibility requirements (see Section 3.1) were removed from the submission file.

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

Survey participant ID	Day of Birth	Month of Birth	Year of Birth	State of Residence
1	31	12	1999	PA
1	30	12	1999	PA
1	15	12	1999	PA
1	31	12	1999	NY
1	30	12	1999	NY
1	15	12	1999	NY

NOTES: Data have been fabricated for this example. Other PII fields not shown as they are the same across all records.
PII – Personally Identifiable Information.

2 Deterministic Linkage Using Unique Identifiers

The deterministic linkage, which was the next step in the linkage process, used only the survey participant and CMS Medicare submission records that included a valid SSN. The algorithm performed two passes on the data, the first pass joining records when all 9-digits of the SSN matched and then for records where the last four digits of the SSN matched. After records had been linked using SSN, the algorithm validated the deterministic links by comparing first name, middle initial, last name, month of birth, day of birth, year of birth, ZIP code of residence, and state of residence. If the ratio of agreeing identifiers to non-missing identifiers was greater than 50% (1st pass using SSN-9) or greater than 2/3 (2nd pass using last 4 of SSN), the linked pair was retained as a deterministic match. In addition to the 2/3's agreement ratio, linked pairs in the 2nd pass were required to have at least 5 non-missing PII variables in agreement to be deemed a deterministic match. Of note, survey participants were excluded from the second pass (i.e., using the last 4-digits of SSN) if they were deterministically linked in the first pass. The collection of records resulting from the deterministic match is referred to as the 'truth source.'

3 Probabilistic Linkage

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

3.1 Blocking

Blocking is a key step in the probabilistic record linkage process. It identifies a smaller set of potential candidate pairs, eliminating the need to compare every single pair in the full comparison space (i.e., the Cartesian product). According to Christen, blocking or indexing, "splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key)."[8] Intuitively developed rules can

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

be used to define the blocking criteria, however, for this linkage, variable values in the data being linked were used to inform the development of a set of blocking passes that efficiently join the datasets together (i.e., multiple, overlapping blocking passes are run, each using a different blocking key). By using these data to create an efficient blocking scheme (or set of blocking passes), a high percentage of true positive links were retained while the number of false positive links were significantly reduced. A supervised machine learning algorithm used the 'truth source' (see Appendix II Section 2) as the validation dataset and the survey participant and CMS Medicare submission records as training data. For more detailed information on the supervised machine learning algorithm used, please refer to "Learning Blocking Schemes for Record Linkage" and "Using supervised machine learning to identify efficient blocking schemes for record linkage".[9] [10]

The machine learning algorithm produced 11 blocking passes to be used in the blocking scheme. Table III provides the PII variables that were assigned to each of the blocking passes and the PII variables that were used to score the potential links in each of the blocking passes. Note, the variables listed in the scoring key are all PII variables not used as a blocking variable.

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

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

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

[10] Campbell, S. R., Resnick, D. M., Cox, C. S., & Mirel, L. B. (2021). Using supervised machine learning to identify efficient blocking schemes for record linkage. Statistical Journal of the IAOS, 37(2), 673–680. https://doi.org/10.3233/SJI-200779.

Key Number	Blocking Key	Scoring Key
10	Last name, month of birth, year of birth	First name, middle initial, sex, day of birth, zip code of residence
11	First name, sex, zip code of residence, state of residence	Middle initial, last name, day of birth, month of birth, year of birth

3.2 Score Pairs

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

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

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

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

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

3.2.1 M and U Probabilities

The M-probability is the probability that the identifiers on a pair of records agree, given that records represent the same person (i.e., the records are a match). M-probabilities were estimated separately within each individual blocking pass and were calculated for each of the identifiers used for scoring (Table III). Within the blocking pass, pairs with agreeing SSN were used to calculate the M-probabilities, as these are assumed to represent the same individual. SSN agreement was defined as having 8 or more digits being the same for pairs with a full 9-digit SSN or the last 4-digits being the same for pairs with only a 4-digit SSN (ex. XXXXX9999). Further, to account for the alternate submission records generated during the creation of the submission files, the "best" agreement was taken for each of the scoring variables among the blocked records for each survey participant ID and CMS Medicare beneficiary ID (see Tables IV and V for example of alternate record summarization). Table IV is an example of how the agreement flags for each of the scoring variables in Blocking pass 1 are created. A value of 1 means the information in the variable is exactly matching, while a 0 means they are not. A value of "." (missing) means the scoring variable is missing for one or both data sources. Table V then represents how the multiple submission records in Table IV are summarized into one record for each survey participant ID and Medicare administrative ID. If any of the identifiers agree across multiple records, they are flagged as agree (i.e., set to 1). The summarized records in Table V are then used to estimate the M-probabilities for each of the specific scoring variables.

Table IV. Example of Agreement Flags Using Blocking Pass 1

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

NOTES: Data have been fabricated for the purposes of this example. PII – Personally Identifiable Information.

[1] Agreement status of 1 = match, 0 = non-match, and . = missing values

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

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

NOTES: Data have been fabricated for the purposes of this example. PII – Personally Identifiable Information.

[1] Agreement status of 1 = match, 0 = non-match, and . = missing values

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

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

The U-probability is the probability that the two values for an identifier from paired records agreed given that they were NOT a match. Similar to the M-probabilities, U-probabilities were calculated only for the PII variables not included in the blocking keys and with the exception of first and last names, were computed within the blocking pass. The U-probabilities were computed using records where non-missing SSNs were not in agreement (defined as having less than 5 matching digits when records had a full 9-digit SSN and less than 4 matching digits for records with a 4-digit SSN). In order to avoid skewing U-probabilities in blocking passes that contained a high percentage of deterministic matches, assumed matches (i.e., records where SSN was not in agreement and had majority of the non-missing PII among scoring variables in agreement) were excluded prior to calculating the U-probabilities. For example, when computing the U-probability for first name in blocking pass 3, record pairs that did not agree on SSN that had a majority (i.e., greater than 50%) of the PII among last name, middle initial, and zip code of residence in agreement were excluded from the assumed non-matches. Even though SSN did not agree, these records were assumed to be probable links given that a majority of the PII between the survey and Medicare submission records agreed.

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

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

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

3.2.2 M and U Probabilities for First and Last Names

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

Because of the large number of unique name values, it was impractical to compute U-probabilities specific to each 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 records in the survey submission file and a simple random sample of 3% of records with non-missing name information from the Medicare submission file (approximately 3.0 million records for first name and 3.1 million records for last name).

Complete name tallies (separately, for first and last names) were then produced for the survey submission file. For each level of "common" name (defined as names appearing more than 100 times) on the survey submission file, 100,000 names were randomly selected from the Medicare submission file 3% sample for comparison. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. For each Jaro-Winkler level, the number of names in agreement of the 100,000 randomly selected Medicare file names were then tallied. A 'catch-all' category was calculated for the remaining "rare" names, based on 5 million name pairs generated by randomly sampling from both the list of rare names on the survey file and from the Medicare submission file 3% sample.[11] [12] [13]

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:

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

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

Agreement weights were only assigned to identifiers that had agreeing values. Similarly, non-agreement weights were only assigned to identifiers that had non-agreeing values. A non-agreement weight was always a negative value and reduced the pair weight score. It is important to note that if the M-probability was smaller than the U-probability (i.e., M<U), the pair score (see Section 3.2.4) was not adjusted according to the agreement/non-agreement weight. Because of the logarithmic function, having a M-probability that is smaller than the U-probability would have an inverse effect on the

[11] 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.

[12] 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.

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

identifier agreement weights. That is, an agreement weight computed using a M-probability that was smaller than the U-probability would produce a negative weight, while the non-agreement weight would be positive. For example, if the M-probability for month of birth was 0.989 and the U-probability was 0.9999 then the agreement and non-agreement weights would be as follows,

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

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

3.2.4 Calculate Pair Weight Scores

In the next step, pair weights were calculated for each record in the blocking pass, which were then used in the probability model. The pair weights were calculated differently for each blocking pass (due to different PII variables contributing to the pair weight), but followed the same general process:

- 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, or M-probability was less than the U-probability: no adjustment made to the pair weight

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

3.3 Probability Modeling

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

- Combine pairs across blocking passes (Pair-weights are specific to each blocking pass and are not comparable)

- Select a “best” record among survey participant IDs that have linked to multiple administrative records.
- Select final matches based on a probability cut-off value (discussed in the following Section 4)

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

1. A pair-weight adjustment was computed (Adj_B) specific to blocking pass, B, by taking the log base 2 of the estimated number of matches (within blocking pass B) divided by the estimated number of non-matches in the blocking pass. For convenience, the estimated number of matches, $\widehat{N_{matches,B}}$, used in the first iteration was set to half of the pairs in the blocking pass (i.e., all pairs generated by the blocking pass specification). The number of non-matches was computed by subtracting the estimated number of matches from the number of pairs (regardless of how likely they are to be matches) in the blocking pass.

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

   Note that in the first iteration, it was assumed that $\widehat{N_{matches,B}}$ = $\widehat{N_{non-matches,B}}$, resulting in $Adj_B = 0$. If, however, in a later iteration, the number of matches was estimated to be, $\widehat{N_{matches,B}}$ = 20,000 (for example), out of the number of pairs, $N_{Pairs,B}$ = 1,000,000, then

$$Adj_B \;= log_2\left(\frac{20{,}000}{1{,}000{,}000 - 20{,}000}\right) \approx -5.61$$

2. The odds of a given pair, *P*, being a match were computed in blocking pass, *B*, by taking 2 to the power of the adjusted pair-weight (sum of pair-weight (*PW*) and Adj_B, the blocking pass pair weight adjustment).

$$Odds_{P,B} = 2^{PW_{P,B}+Adj,B}$$

   Continuing with the example from Step 1…
   if for Pair 1 of blocking pass B, the pair-weight is 8.4, then $Odds_{1,B} = 2^{(8.4+\,-5.61)}$ ≈6.9
   if for Pair 2 of blocking pass B, the pair-weight is -2.5, then $Odds_{2,B} =$ $2^{(-2.5+\,-5.61)}$ ≈0.0036
   …and this continues for the remaining $N_{Pairs,B}$ pairs of the blocking pass

3. Each record pair had a match probability estimated using the odds. This was accomplished by taking the odds for pair, P, in blocking pass, B, and dividing by the (Odds+1).

$$P_{EM,P,B}(Match) = \left(\frac{Odds_{P,B}}{Odds_{P,B} + 1}\right)$$

   Continuing with the example…

For Pair 1 in blocking pass B, $P_{EM,P,B}(Match) = \left(\frac{6.9}{6.9+1}\right) \approx 0.87$

For Pair 2 in blocking pass B, $P_{EM,P,B}(Match) = (\frac{0.0036}{0.0036+1}) \approx 0.0036$

…and this continues for the remaining $N_{Pairs,B}$ pairs of the blocking pass.

4. The new number of matches in blocking pass were estimated. This was done by summing each of the estimated probabilities in the block.

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

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

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

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 $N_{\widehat{matches},B}$ to be estimated. These estimated probabilities were then used to select the final matches, as described below in Section 4.

3.4 Adjustment for SSN Agreement

Up to this point, every pair generated through the probabilistic routine was assigned a value that estimates its probability of being a match. However, this estimate did not take SSN agreement into account. This was conducted as a separate step because for the other comparison variables, M- and U-probabilities were estimated based on probable matches or non-matches that were determined based on SSN agreement, and clearly this was infeasible for SSN itself.[14]

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

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

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

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

When the last four digits of SSN did not agree:

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

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

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

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

4.1 Estimating Linkage Error to Determine Probability Cut-off Value

Subsequent to performing the record linkage analysis an error analysis was performed. There are two 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.

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

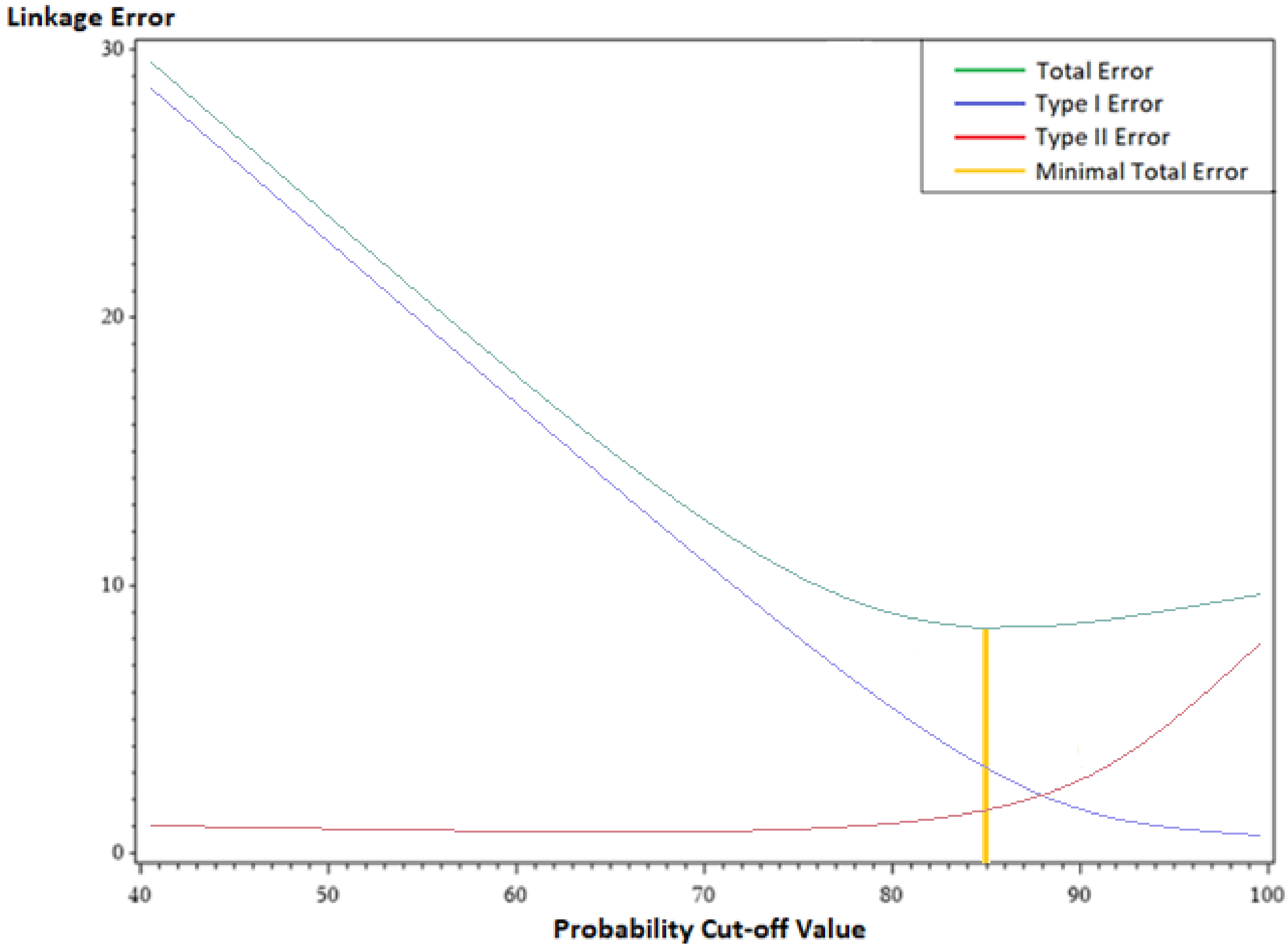
To measure Type II error, the truth source comprised of all matches identified in the deterministic linkage was used. Recall, the truth source contains records with full nine-digit SSN agreement (step 1) or with the last four digits of SSN in agreement (step 2). Potential deterministic matches were then validated using the available PII (see, Appendix I section 2). It was expected that this truth source had only a few exceptional pairs that were not true matches. For the probabilistic records, Type II error was estimated as the percentage of the truth source records that were not returned as links by the probabilistic method. Similarly to the computation of Type I error, an adjustment was made to the Type II error since some links having agreeing SSNs were being linked deterministically even if they were not returned by the probabilistic approach. For example, say that the probabilistic approach was able to return 97% of true matches as links. If only a probabilistic linkage was conducted, the Type II error would then be 3%. However, among the 3% not linked probabilistically, some pairs could be linked deterministically. If the deterministic linkage rate is 50% (and if we assume the same rate among the non-linked pairs), then the Type II error rate can be estimated as 0.5*(1 – 0.97) = 0.015 or 1.5%.

4.2 Set Probability Cut-off Value

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

(Illustrative schematic not based on actual values)



4.3 Select Links Using Probability Cut-off Value

The final step in the linkage algorithm was to determine links, which were record pairs inferred to be matches. Links were pairs where the $Probvalid_{SSN_{Adj}}$ exceeded the probability cut-off value (from Section 4.2). Further, the 'best' record pair (i.e., highest $Probvalid_{SSN_{Adj}}$) among the records that exceeded the probability cut-off value was selected for each survey participant. All record pairs with an adjusted probability value that fell below the probability cut-off value were not linked.

4.4 Computed Error Rates of Selected Links

Final error rates were computed for selected links (described in Section 4.3). Table VI provides the total number of selected links, the number of total links identified through deterministic and probabilistic methods, and the Type I and Type II error rates for the household survey linkage. Because the links were selected using the SSN adjusted probability (described in Section 4.1), the overall Type I error rate was computed using the estimated match probabilities rather than using SSN agreement. For the probabilistic links, the estimated match probabilities represented the probability that the survey participant record was a match to the Medicare administrative record. In other words, if a link had an estimated probability of 0.98, then it was understood that there was a 98% chance this was a match. To estimate the Type I error rate for the probabilistic links, the chance that a link is not a match was summed (i.e., $\sum 1 - Probvalid_{SSN_{Adj}}$) and then divided by the total number of probabilistic records. The method to measure the overall Type II error remained unchanged (see Section 4.1).

Table VI. Algorithm Results for Total Selected Links

	Probability Cut-off Value	**Total Selected Links**	**Deterministic Matches**	**Probabilistic Links**	**Est Incorrect (Type I)**	**Est Not Found (Type II)**
Household surveys	0.85	247,082	165,030	82,052	0.03%	0.73%

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

The data provided on the 1999–2021 NHIS, 1999–2018 and 2017–March 2020 Pre-Pandemic NHANES, and NHANES III 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).

Note: At this time the linked CMS Medicare data files are only available for research use through the NCHS RDC Network. 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 or SEQN) 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 or SEQN is provided below.

1 National Health Interview Survey (NHIS), 1999–2021

1.1 NHIS, 1999–2003

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.

*The person identifier was called PX in the 1999–2003 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).

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)
```

R example:

```
# Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%04d", df$SRVY_YR), sprintf("%06d", df$HHX),sprintf("%02d",
df$FMX),sprintf("%02d", df$PX))
```

1.2 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)
```

R example:

```
# Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%04d", df$SRVY_YR), sprintf("%06d", df$HHX),sprintf("%02d",
df$FMX),sprintf("%02d", df$FPX))
```

1.3 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)
```

R example:

```
# Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%04d", df$SRVY_YR), sprintf("%06d", df$HHX),sprintf("%02d",
df$FMX),sprintf("%02d", df$FPX))
```

1.4 NHIS, 2019–2021

Variable	Public-use Location	Length	Description
SRVY_YR	3-6	4	Year of interview
HHX	7-13	7	Household number
RECTYPE	1-2	2	Record type

Note: The NHIS public-use files since the 2019 redesign do not contain a Person Number variable. To merge multiple NHIS public-use files, follow instructions provided in NHIS documentation. To merge to the linked data, concatenate the above variables from the Sample Adult file (so RECTYPE=10 for all Sample Adults) to get the unique person identifier.

SAS example:

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

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

```
egen PUBLICID = concat(SRVY_YR HHX RECTYPE)
```

R example:

```
df$PUBLICID <-paste0(df$SRVY_YR, df$HHX, df$RECTYPE )
```

2 National Health and Nutrition Examination Survey (NHANES), 1999–2018 and 2017–March 2020 Pre-Pandemic Data

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.