The Linkage of the 2021 National Ambulatory Medical Care Survey Health Center Component to 2020–2022 Transformed Medicaid Statistical Information System Data: Linkage Methodology and Analytic Considerations

Data Release Date: September 29, 2025

Document Version Date: September 29, 2025

Division of Analysis and Epidemiology,

National Center for Health Statistics,

Centers for Disease Control and Prevention

Suggested Citation: National Center for Health Statistics. Division of Analysis and Epidemiology. *The Linkage of the 2021 National Ambulatory Medical Care Survey Health Center Component to 2020-2022 Transformed Medicaid Statistical Information System (T-MSIS) Data: Linking Methodology and Analytic Considerations*, September 2025. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/linked-data/namcs/index.html.

Contents

List of Acronyms

CHIP, Children's Health Insurance Program

CMS, Centers for Medicare & Medicaid Services

DE, Demographic and Eligibility

DOB, date of birth

DQ, data quality

ED, emergency department

EHR, electronic health record

EM, expectation-maximization

ERB, Ethics Review Board

FFS, fee-for-service

FPL, Federal Poverty Level

FQHC, Federally Qualified Health Center

HC, Health Center

HUD, U.S Department of Housing and Urban Development

IP, inpatient services

LT, long-term care services

M-CHIP, Medicaid expansion Children's Health Insurance Program

NAMCS, National Ambulatory Medical Care Survey

NCHS, National Center for Health Statistics

NDC, National Drug Code

NDI, National Death Index

OP, outpatient

OT, Other services

PH, Public Housing

PII, personally identifiable information

PW, pair weight

RDC, Research Data Center

ResDAC, Research Data Assistance Center

RX, Prescription drug services

S-CHIP, State Children's Health Insurance Program

SPA, State Plan Amendment

SSDI, Social Security Disability Insurance

SSI, Supplemental Security Income

SSN, Social Security number

TAF, Transformed Medicaid Statistical Information System Analytic File

TANF, Temporary Assistance for Needy Families

T-MSIS, Transformed Medicaid Statistical Information System

UB-04, uniform billing form

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 nation. NCHS collects data from many sources of health information including birth and death records, population-based surveys, and medical records. In 2021, National Ambulatory Medical Care Survey (NAMCS) began collecting electronic health records (EHRs) for ambulatory care visits that took place in health centers, known as the 2021 NAMCS Health Center (HC) Component. Even though the 2021 NAMCS HC Component is a provider survey (i.e., health centers are the sampling unit) it collects patient level personally identifiable information (PII), which enable data linkages.

Through its Data Linkage Program, NCHS has been able to expand the analytic utility of the 2021 NAMCS HC Component data by linking the 2021 NAMCS HC patient encounter data with Medicaid and Children's Health Insurance Program (CHIP) enrollment and utilization data extracted from the Transformed Medicaid Statistical Information System (T-MSIS) collected by the Centers for Medicare & Medicaid Services (CMS). This report will describe the linkage of the 2021 NAMCS HC Component to 2020-2022 T-MSIS records. The linkage of the 2021 NAMCS HC Component patient data with 2020-2022 Medicaid and CHIP enrollment and utilization data creates a new data resource that can support a wide array of public health surveillance and policy evaluation studies based on a more complete picture of health care services utilization and patient health outcomes.

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

The linkage of the 2021 NAMCS HC Component and the 2020-2022 T-MSIS data was performed at NCHS through contract #HHS75D30123A17667 by NORC at the University of Chicago with funding from the Department of Health and Human Services' Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) under a project titled "Linking Federally Qualified Health Center EHRs and Medicaid Data for Increased Data Capacity to Understand Maternal Health Care.".

2 Background on Linked Files

2.1 National Ambulatory Medical Care Survey (NAMCS) Health Care (HC) Component

The National Ambulatory Medical Care Survey (NAMCS), administered by NCHS, is a national survey designed to meet the need for objective, reliable information about the provision and use of ambulatory medical care services in the United States. First fielded in 1973, NAMCS is one of the NCHS National Healthcare Surveys, a family of provider-based surveys, covering a broad spectrum of health care settings. In 2012, NCHS added a separate national sample of community health centers to NAMCS in order to produce nationally representative estimates on health center provided ambulatory care services. Beginning in 2021, NCHS developed a NAMCS Health Center (HC) Component sampling frame to focus on the collection of Electronic Health Records (EHR) for all patient visits within a sampled HC. Sampled health centers include both Federally Qualified Health Centers (FQHCs) and FQHC look-alikes in

the 50 U.S. states and the District of Columbia, which provide ambulatory (or direct outpatient) care to the public and use an EHR system in one or more of their delivery sites.[1]

Participating health centers were asked to submit EHR data for all patient encounters in calendar year 2021. These data included patient identifiers such as name, address, and social security number (SSN) when it is available; date of visit; diagnoses and services provided or ordered during the visit; reason for visit; and clinical notes. In calendar year 2021, 29 of the 111 sampled health centers submitted data on 3,543,927 patient visits from January 1, 2021 to December 31, 2021, for an unweighted response rate of 26.1% and a weighted response rate of 26.8%. The 2021 NAMCS HC Component data are weighted and can be used to produce nationally representative estimates for visits at FQHCs.[2] NCHS did not release a public use file for the 2021 NAMCS HC Component, but rather made the data available via the NCHS Research Data Center (RDC) Network. More information about accessing the restricted-use 2021 NAMCS HC Component data files can be found at the NCHS Research Data Center.

2.2 Centers for Medicare & Medicaid Services (CMS) Medicaid and CHIP Programs

2.2.1 Medicaid

Medicaid is a health insurance program administered by states, according to federal requirements, and funded jointly by states and the federal government. Medicaid provides health care coverage to certain populations, such as:

- Pregnant women
- Children under 19 years of age
- Eligible low-income adults
- Elderly and disabled individuals

State variation in the implementation of Medicaid, such as setting eligibility requirements, benefits, provider payments, and service delivery models, leads to variation in enrollment and spending across states. More information on Medicaid enrollment data and eligibility is available at https://www.medicaid.gov/medicaid.

2.2.2 Children's Health Insurance Program (CHIP)

CHIP is a health insurance program administered by states, according to federal requirements, and funded jointly by states and the federal government. CHIP provides coverage to eligible children in families with incomes that exceed Medicaid program eligibility requirements but are considered too low to afford private insurance, such as:

- Children under 19 years of age
- Pregnant women (in some states)

States may choose to offer CHIP coverage through the state's Medicaid program (M-CHIP) or as a separate state CHIP program (S-CHIP) or a combination of both. Because states may design their own

[1] More detailed information about the 2021 NAMCS sampling procedures are available at https://www.cdc.gov/nchs/data/series/sr_02/sr02-203.pdf.
[2] More detailed information on producing weighted analyses based on the 2021 NAMCS HC Component data is available at: https://www.cdc.gov/rdc/data/b1/2021-NAMCS-HCC-RDC-Data-Dictionary-508.pdf.

CHIP program, within federal guidelines, benefits vary by state and CHIP program type. More information on CHIP eligibility and enrollment and state program information is available at https://www.medicaid.gov/chip.

2.3 Transformed Medicaid Statistical Information System (T-MSIS) Data

T-MSIS collects Medicaid and CHIP data from U.S. states, territories, and the District of Columbia. States submit standardized data files to CMS on a monthly basis that include information on Medicaid and CHIP eligibility, enrollment, service use, and payments. State T-MSIS data submissions are derived from administrative data that are created for program administration purposes, such as enrolling individuals, adjudicating and paying healthcare claims, certifying and enrolling providers, assuring fiscal integrity, assessing quality, and performing other program management functions. CMS publishes a Medicaid and CHIP data quality assessment resource known as the Medicaid.gov Data Quality (DQ) Atlas which provides information on the quality of Medicaid and CHIP T-MSIS data by topic area and state.

2.3.1 T-MSIS Analytic Files (TAFs)

To facilitate analysis of Medicaid and CHIP data, CMS creates annual T-MSIS Analytic Files (TAFs). The 2021 NAMCS HC Component linkage with 2020-2022 T-MSIS data included the following TAFs:

- Demographic and Eligibility (DE) file provides demographic, program eligibility, and enrollment information on each person who was enrolled for at least one day in the calendar year in Medicaid and/or CHIP.
- Inpatient Hospital (IP) file includes records for inpatient hospital services for Medicaid and CHIP enrollees during the calendar year.
- Long-Term Care (LT) file includes records for institutional long-term care services for Medicaid and CHIP enrollees during the calendar year.
- Pharmacy (RX) file includes records for prescribed drugs, supplies, and other items provided by a free-standing pharmacy, either directly to a Medicaid or CHIP enrollee or to a long-term facility for the enrollee's use.
- Other Services (OT) file includes records for all other community-based health care services for Medicaid and CHIP enrollees not reported in the IP, LT, and RX files.

More detailed information about the TAFs is available in Appendix II and at www.resdac.org and www.ccwdata.org. In the TAF claims files (IP, LT, RX, OT), original state submitted claims, voids, credits, and debits are resolved to create final action claims[3].

[3] For more information on final action claims please visit: https://www.medicaid.gov/dq-atlas/landing/briefs and download the Final Action Status in TMSIS claims analytic brief. (NOTE: Special circumstances may pertain to linked TAF claims data submitted by Illinois – See TAF Technical Guidance: How to Use Illinois Claims Data for more information). The T-MSIS variable SUBMTG_STATE_CD can be used to identify Illinois claims.

3 Linkage Methodology

3.1 Linkage Eligibility Determination

The linkage of 2021 NAMCS HC Component patient records to 2020-2022 T-MSIS data was conducted through an agreement between NCHS and CMS. Approval for the linkage of NAMCS HC Component patient data with T-MSIS data was provided by NCHS' Research Ethics Review Board (ERB).[4]

Linkage was attempted only for 2021 NAMCS HC Component patient records that had at least two of the following three identifiers present:

- valid SSN[5]
- valid date of birth (month, day, and year)[6]
- valid name (first, middle initial, and last)[7]

For example, if the PII on the 2021 NAMCS HC Component patient record had no SSN, a full name, and only the year of birth, the record would be considered ineligible for linkage, as only one of the criteria (i.e., valid name) was met.

The variable ELIGSTAT, included on the linked 2021 NAMCS HC Component – T-MSIS match file, provides the linkage eligibility status for each NAMCS HC Component patient record. ELIGSTAT values include 0 (ineligible) or 1 (eligible). The 2021 NAMCS HC Component included 726,491 (99.9%) patients who were determined to be eligible for linkage with CMS T-MSIS administrative data. Note that linkage eligibility is distinct from program eligibility, which defines whether a person meets the eligibility criteria for a specific government-administered or funded program.

3.2 Linkage Overview

This section outlines steps that were used to link the 2021 NAMCS HC Component data to the 2020-2022 T-MSIS enrollment data. The linkage was conducted at a patient level using patient identifiers collected from health center submitted patient visit records. For more detailed information on linkage methodology, see Appendix I.

Linkage-eligible 2021 NAMCS HC Component patient records were linked to the CMS T-MSIS enrollment database using the following identifiers: SSN, first name, last name, middle initial, month of birth, day of birth, year of birth, 5-digit ZIP code of residence, state of residence, and sex.

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

[5] 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-xx0000), and digits are not consecutive (ex. 012345678). Additionally, special SSN values (i.e., 123-123-1234, 111-22-3333, 010-010- 0101, 001-01-0001, etc.) were changed to missing. Four-digit SSN is considered valid if: 4-digits in length, containing only numbers, and is between 0001 and 9999.

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

[7] 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.

The 2021 NAMCS HC Component patient records and the CMS T-MSIS enrollment database were linked using both deterministic and probabilistic approaches and were conducted separately for males and females[8].

1. Deterministic linkage joined records on exact SSN and validated links by comparing other identifying fields (i.e., first name, last name, day of birth, etc.)
2. Probabilistic linkage identified likely matches, or links, between all records. All records were probabilistically linked[9] and scored as follows:
    A. Formed pairs via blocking
    B. Scored all pairs
    C. Modeled probability – assigned estimated probability that pairs are matches
3. Pairs were selected that were believed to represent the same individual between data sources (i.e., they are a match).
    A. Deterministic matches (from step 1) were assigned a match probability of 1
    B. Records selected from the probabilistic match (step 2) were assigned the modeled match probability.

A score threshold was established for determining which 2021 NAMCS HC Component patients were considered linked to 2020-2022 T-MSIS data. For each 2021 NAMCS HC Component patient-level record that was determined to be linked, CMS extracted the associated 2020-2022 TAF file records (DE, IP, LT, RX, OT) and sent the data to NCHS following secure data transfer procedures. Table 1 highlights the linkage results.

[8] Because first names are commonly associated with a person's sex, conducting the linkage separately for males and females helps to ensure independence and more appropriate weighting of name comparisons. Additionally, multiple part first and last names are more likely to be associated with females, which are handled differently when creating the linkage submission file. See Appendix I, Section 1 for additional information on the alternate record generation process for multiple part names.

[9] The probabilistic linkage methodology used is based on Fellegi, I. P., and Sunter, A B. (1969), "A Theory for Record Linkage," JASA 40 1183-1210.

Table 1. Linked 2021 NAMCS HC Component – 2020-2022 T-MSIS Records: Sample Sizes and Percent Linked, by Age and Sex

	Sample Size			Percent Linked	
	Total Sample	Eligible for Linkage[2]	Linked to T-MSIS Claims Data[3]	Total Sample[4]	Eligible Sample[5]
2021 NAMCS HC Component					
Age[1]					
0-17	185,995	185,990	149,029	80.1	80.1
18-44	271,369	271,360	150,894	55.6	55.6
45-64	184,114	184,111	83,673	45.4	45.4
65 and over	85,031	85,025	31,264	36.8	36.8
Not Calculated	848	5	2	0.2	40
Total	727,357	726,491	414,862	57.0	57.1
Sex					
Male	309,847	309,832	168,073	54.2	54.2
Female	414,202	414,194	245,937	59.4	59.4
Missing	3,308	2,465	852	25.8	34.6
Total	727,357	726,491	414,862	57.0	57.1

NOTES: Data are presented at patient level.
[1] Age is as of final health center visit (date of last known contact). Age is calculated by subtracting patient date of birth (DOB) from the final visit date. When more than one DOB was present, the minimum of the non-missing DOB was selected.
[2] Eligibility for linkage is based upon having sufficient PII in at least two of three data element groups: SSN, name, and date of birth.
[3] This group includes linkage-eligible patients who linked to T-MSIS data as of final health center visit (date of last known contact).
[4] This percentage is calculated by dividing the number of linked patients by the number of patients in the total sample.
[5] This percentage is calculated by dividing the number of linked patients by the total number of linkage-eligible patients.

4 Analytic Considerations

4.1 Analytic Considerations for 2021 NAMCS HC Component Data

4.1.1 2021 NAMCS HC Component Restricted-Use Files (RUF)

The 2021 NAMCS HC Component restricted-use survey data are made available for analysis through the NCHS RDC network. For more information about obtaining access to the 2021 NAMCS HC Component restricted-use files (RUFs) see Section 5.0. The 2021 NAMCS HC Component RUFs are organized as relational data tables organized by Visits, Patients, Conditions, Procedures and Weights. For more information about the specific variables and the observational unit for each data table please see the 2021 NAMCS HC Component data RUF documentation.

During the 2021 NAMCS data reporting period some health centers did not provide certain data elements for any of their reported visits. NCHS has provided detailed analytic guidance on how users should adjust their analysis of 2021 NAMCS HC Component data to account for missing data. More detailed information on missing data adjustment is available from the 2021 NAMCS HC Component data RUF documentation.

4.1.2 Using 2021 NAMCS HC Component Visit Weights

The NCHS Division of Health Care Statistics (DHCS) has produced visit weights that can be used to produce nationally representative estimates of ambulatory care visits occurring in FQHCs. For more detailed information regarding producing weighted estimates with 2021 NAMCS HC Component data, please see the 2021 NAMCS HC Component data RUF documentation.

The linkage between the 2021 NAMCS HC Component data and 2020-2022 T-MSIS data was conducted at a patient level using patient identifiers collected from health center submitted patient visit records. The patient identifiers collected from the visit records were used to link T-MSIS enrollment and utilization data covering the calendar years 2020 through 2022. For 2021 NAMCS HC Component patients with linked 2020-2022 T-MSIS data, it will be possible for analysts to align the month of the 2021 NAMCS HC Component patient visit with 2021 T-MSIS enrollment periods and apply the 2021 NAMCS HC Component visit weights

Because the linked T-MSIS data cover the time period from 2020-2022, 2021 NAMCS HC Component patients may have Medicaid enrollment and utilization information for the calendar year prior to the 2021 NAMCS data reporting period and the calendar year after. Although DHCS has developed visit level weights for use with the 2021 NAMCS HC Component visit data, patient level weights have not been created. Analysts wishing to analyze the linked 2020-2022 T-MSIS data at the patient level will not be able to perform weighted analyses.

4.2 Analytic Considerations for T-MSIS Data

This section summarizes some key analytic considerations when analyzing T-MSIS data. It is not an exhaustive list of the analytic issues that analysts may encounter. Analysts are encouraged to consult the CMS T-MSIS Data Guide and the Medicaid Data Quality Atlas for additional technical information regarding T-MSIS data. Additional T-MSIS TAF technical guidance is also available at www.resdac.org[10] and www.ccwdata.org.

4.2.1 State Differences in Medicaid and CHIP

Medicaid and CHIP programs vary at the state level. Program eligibility, covered services, managed care enrollment, provider reimbursement and other program factors vary from state to state (see Section 2.2.1 and 2.2.2 for more information on the Medicaid and CHIP programs). Furthermore, there may be variation in the quality of T-MSIS data across states and within a state over time.

The T-MSIS data element SUBMTG_STATE_CD, included in all linked TAF files (DE, IP, LT, RX, OT) should be requested in the analyst's RDC application if the analyst intends to incorporate state program characteristics in their analyses. However, although analysts may incorporate state level Medicaid and CHIP program characteristics in their analyses, due to NCHS disclosure concerns, it may not be possible to publish state-level estimates based on linked 2021 NAMCS HC Component – 2020-2022 T-MSIS data. Requests to conduct state-level analyses will be assessed for disclosure risk through the NCHS RDC application review process.

[10] Detailed CMS guidance on analyzing TAFs is located at: https://resdac.org/sites/datadocumentation.resdac.org/files/2021-01/TAF_TechGuide_DE_File.pdf and https://resdac.org/sites/datadocumentation.resdac.org/files/2021-01/TAF_TechGuide_Claims_Files.pdf

4.2.2 Determining Medicaid Program Enrollment

An individual's eligibility for Medicaid may change over time. As an individual's eligibility changes, they may enroll and disenroll in Medicaid throughout the calendar year. The DE TAF variables MDCD_ENRLMT_DAYS_01 through MDCD_ENRLMT_DAYS 12 provide the number of days of Medicaid enrollment in each month of the year and the variable MDCD_ENRLMT_DAYS_YR provides the total number of days of Medicaid enrollment for the calendar year. The variable FILE_YEAR4 is used to identify the calendar year (2020-2022) of interest.

4.2.3 Determining CHIP Program Enrollment

An individual's eligibility for CHIP may change over time. As an individual's eligibility changes, they may enroll and disenroll in CHIP throughout the calendar year. The DE TAF file variables CHIP_ENRLMT_DAYS_01 through CHIP_ENRLMT_DAYS 12 provide the number of days of CHIP enrollment in each month of the year and the variable CHIP_ENRLMT_DAYS_YR provides the total number of days of CHIP enrollment for the calendar year. The variable FILE_YEAR4 is used to identify the calendar year (2020-2022) of interest. CHIP enrollment program type (S-CHIP or M-CHIP) can be identified by using the T-MSIS variables CHIP_CD_1 to CHIP_CD_12 on the DE TAF.

4.2.4 Identifying Medicaid Restricted Benefit Enrollees

States have the option to limit certain Medicaid enrollees to a set of restricted benefits including limiting Medicaid covered services to only family planning, pregnancy care, or substance use disorder treatment. Information on specific restricted benefits enrollment is available by month on the DE TAF in variables RSTRCTD_BNFTS_CD_01 through RSTRCTD_BNFTS_CD_12. Analysts should consider whether it is appropriate to remove restricted benefit enrollees from their linked 2021 NAMCS HC Component – 2020-2022 T-MSIS analyses.

4.2.5 Identifying Dually Eligible Individuals

Dually eligible individuals are individuals enrolled in both Medicare and Medicaid. For these individuals, Medicare is the primary payer for services it covers, while Medicaid may cover Medicare cost-sharing (e.g. premiums, deductibles, and coinsurance) and provide additional benefits not covered by Medicare. Medicaid health care claims for dually eligible individuals contain information about the Medicaid covered portion of the provided health care service but do not contain information about Medicare covered services. Information on dual program eligibility is available by month on the DE TAF in variables DUAL_ELGBL_01 through DUAL_ELGBL_12. Analysts using linked 2021 NAMCS HC Component – 2020-2022 T-MSIS data files should consider if and how they want to include dually eligible individuals in their analyses.

4.2.6 Managed Care

Medicaid and CHIP services may be delivered through both fee-for-service (FFS) and managed care programs. Medicaid managed care is a health care delivery system in which states contract with managed care organizations to provide Medicaid benefits to enrollees, typically for a fixed per member per month payment. Medicaid managed care payment information is not available in the linked 2021 NAMCS HC Component – 2020-2022 T-MSIS data files. However, the linked 2020-2022 TAF claims files (IP, LT, RX, OT) do include encounter records that represent the services provided under both fee-for

service and managed care programs. Information about Medicaid managed care program enrollment is available by month on the DE TAF in variables MC_PLAN_TYPE_CD_01 through MC_PLAN_TYPE_CD_12.

4.2.7 Waiver and Demonstration Reporting

Individuals can be enrolled in various state waivers. The DE TAF does not include information on waiver enrollment, but the linked 2020-2022 TAF claims files (IP, LT, RX, OT) include data elements that can be used to identify services provided under waivers. T-MSIS data variable WVR_TYPE_CD identifies the type of waiver under which a service was provided, and T-MSIS data variable WVR_ID is the state-assigned identifier for the waiver. Analysts interested in analyzing specific state-based waivers should use the information provided in WVR_TYPE_CD and WVR_ID as well as submitting state code, SUBMTG_STATE_CD.

4.2.8 Service Tracking Claims Records

Most claims are submitted for individual enrollees, but states may submit a small percentage of claims records, known as service tracking claims, for a group of enrollees. Use of these types of claims varies by state. Because service tracking claims cannot be linked to an individual, they have been excluded from the linked 2021 NAMCS HC Component – 2020-2022 T-MSIS files.

4.2.9 Missing Enrollment Data and Dummy Enrollment Records

There are instances in which there are valid claim records from the linked 2020-2022 TAF claims files (IP, LT, RX, OT) for an enrollee, but there is no associated state-reported enrollment record. CMS has created 'dummy' enrollment records for these enrollees in the DE TAF file. DE 'dummy' records can be identified using the variable MISG_ELGBLTY_DATA_IND, code value = 1. 'Dummy' enrollment records typically do not include enrollee demographic information.

4.2.10 Header and Line-Item Claims Records

T-MSIS claims include both header records (which provide a summary of services provided) and line-item records (which contain the detail on services provided). The sum of payment amounts in line-item records may not equal the total payment amount on header records. Also, some line-item records may show $0 paid amounts. The Medicaid DQ Atlas includes an analysis of payment consistency between header and line-item claims records for the four claims file types. The TAF claims files (IP, LT, RX, OT) include a variable (PYMT_LVL_IND) that indicates whether the Medicaid claim payment was made at the header or line-item level.

4.2.11 Mother and Newborn Claims Records

States use different methods to report labor and delivery services provided to women and their newborns. Some providers may report services provided to the newborn using the mother's Medicaid ID. Other providers may report services provided to the mother using the newborn's Medicaid ID. CMS has published analytic guidance to assist analysts in understanding the frequency of shared Medicaid IDs across states and years. [11] Analysts interested in examining state level differences should request the variable SUBMTG_STATE_CD in their RDC application.

[11] For more information on shared Medicaid IDs please visit: https://www.medicaid.gov/dq-atlas/landing/briefs and download the Use of the Same Medicaid ID Number for Live-Birth Delivery and Newborn Services in 2022 analytic brief.

4.2.12 Multiple Claims with the Same Service Date

Due to the manner in which health care claims are submitted for reimbursement for certain types of services, there can be multiple claims for the same enrollee with the same date of service. These are not errors or data anomalies, but instead distinct services or portions of a provided service that were billed separately.

4.3 Analytic Considerations for Linked 2021 NAMCS HC Component – 2020-2022 T-MSIS Data Files

This section summarizes some key analytic issues for users of the linked 2021 NAMCS HC Component - 2020-2022 T-MSIS data files. It is not an exhaustive list of the analytic issues that analysts may encounter while using the linked 2021 NAMCS HC Component- 2020-2022 T-MSIS data. Analysts are encouraged to consult the CMS T-MSIS Data Guide and the Medicaid Data Quality Atlas for additional technical information regarding CMS T-MSIS data files. Additional TAF technical guidance is also available at www.resdac.org and www.ccwdata.org.

4.3.1 T-MSIS Match File

NCHS produces a linked T-MSIS Match file that can be used to identify which of the 2021 NAMCS HC Component patients were eligible for linkage and linked to a 2020-2022 T-MSIS DE TAF record. This file contains one record for each 2021 NAMCS HC Component patient ID and includes the variables ELIGSTAT, PROBVALID, and TMSIS_MATCH_STATUS.

The variable ELIGSTAT should be used to determine linkage eligibility (Section 3.1). 2021 NAMCS HC Component patient IDs with an ELIGSTAT value of 1 were considered eligible for linkage to the T-MSIS demographic and eligibility records.

Data linkages include some uncertainty over which pairs represent true matches. An estimated probability of match validity (PROBVALID) was computed for each candidate pair and compared against a probabilistic cut-off value to determine which pairs were links (an inferred match). For additional discussion on how PROBVALID was estimated, see Appendix I, Sections 3.3 and 3.4. NCHS used a probabilistic cut-off value which minimized the total estimated counts of Type I error (false positive links – identified as enrolled in Medicaid but actually are not) and Type II error (false negative links – identified as not enrolled in Medicaid but actually are).

In the 2021 NAMCS HC Component- 2020-2022 T-MSIS linkage, NCHS used a probabilistic cut-off value of 0.85 to determine final match status. Candidate pairs with a PROBVALID that exceeded the probabilistic cut-off (i.e., PROBVALID>0.85) were deemed a link. The estimated Type I error was <0.1% and the Type II error was 0.3%**.** For additional discussion on probabilistic cut-off determination and record selection, please see Appendix I, Section 4. For some analyses, it may be desirable to reduce the Type I error. To do this, analysts should increase the probability cut-off value (to a value closer to 1.0). Of note, the PROBVALID cannot be decreased from 0.85. To change the NCHS link acceptance cut-off value, analysts should request the variable PROBVALID in their RDC application.

The variable TMSIS_MATCH_STATUS should be used to identify which of the 2021 NAMCS HC Component patients were successfully linked to a T-MSIS record during the 2020-2022 T-MSIS linkage

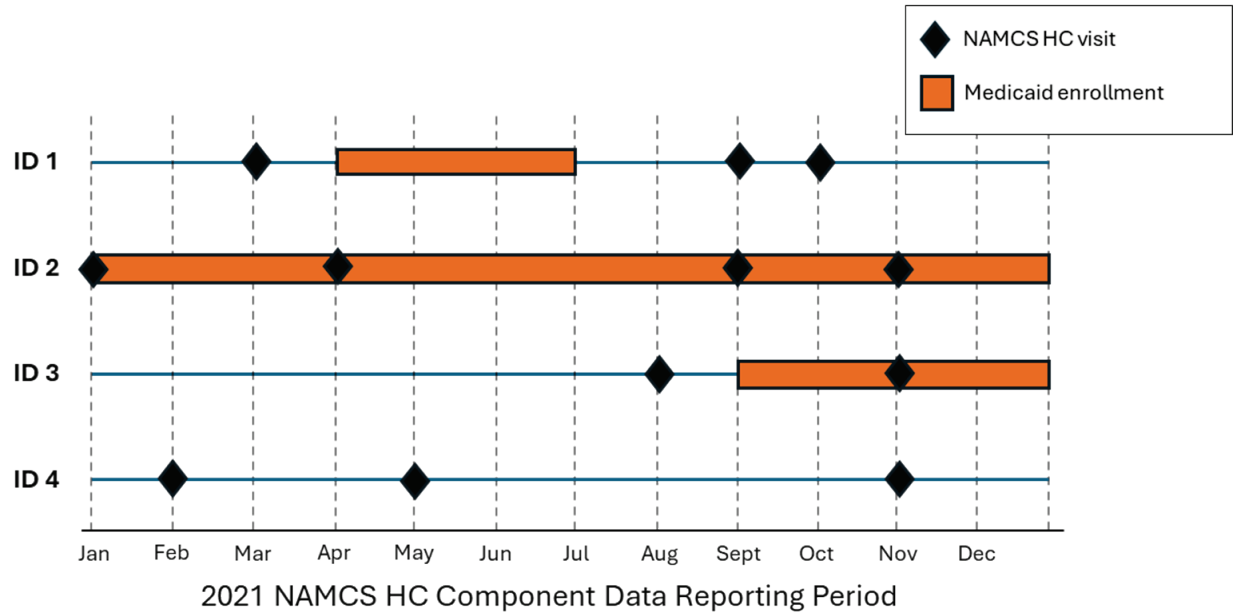
period. 2021 NAMCS HC Component patient IDs with a TMSIS_MATCH_STATUS value of 1 were successfully linked to a 2020-2022 T-MSIS DE TAF record.

4.3.2 Temporal Alignment of Medicaid or CHIP Enrollment with 2021 NAMCS HC Component Patient Visit Data

To identify whether the 2021 NAMCS HC Component patient was enrolled in Medicaid or CHIP during, before, or after a specific health center visit, analysts can compare the month and year of the 2021 NAMCS HC Component patient visit (VISIT_START_DATETIME) with Medicaid enrollment variables that provide a count of the number of days for each month of Medicaid enrollment (MDCD_ENRLMT_DAYS_01 Thru 12). Similar variables are available for CHIP enrollment (CHIP_ENRLMT_DAYS_01 THRU 12). Analysts should request the 2021 NAMCS HC Component variable (VISIT_START_DATETIME) in their RDC application to infer whether a 2021 NAMCS HC Component patient visit (VISIT_START_DATETIME) occurred during a period of Medicaid or CHIP enrollment.

Figure 1 below depicts potential temporal alignment scenarios for 2021 NAMCS HC Component patient visit data and monthly 2021 Medicaid enrollment data for four hypothetical patients noted as patient ID1 through patient ID4. In each timeline, the diamond represents the month during which the 2021 NAMCS HC Component patient visit occurred, and the orange bar represents the month(s) during which the patient was enrolled in Medicaid or CHIP. For example, Patient ID1 was enrolled in Medicaid during the months of April, May, and June in 2021 but was not enrolled in July 2021. The orange bar of Patient ID 1 spans only the months April through June and stops at July.

Figure 1. Temporal alignment of 2021 NAMCS HC Component patient visit data linked to 2021 Medicaid enrollment data.



Sources: 2021 NAMCS HC Component patient data linked to 2020-2022 T-MSIS administrative data.
Note: T-MSIS is the Transformed Medicaid Statistical Information System.

The examples shown in Figure 1 are as follows,

- Patient ID1 was enrolled in Medicaid during the 2021 NAMCS HC Component data reporting period, but their Medicaid enrollment period was not concurrent with any of the months in which Patient ID1 had a 2021 NAMCS HC Component reported health center visit.
- Patient ID2 was enrolled in Medicaid for all months in calendar year 2021 and had four health center visits during this period.
- Patient ID3 was enrolled in Medicaid during the month of their November 2021 health center visit but was not enrolled in Medicaid during their August 2021 health center visit.
- Patient ID4 had three health center visits but was not enrolled in Medicaid at the time of any of their 2021 NAMCS HC Component reported health center visits.

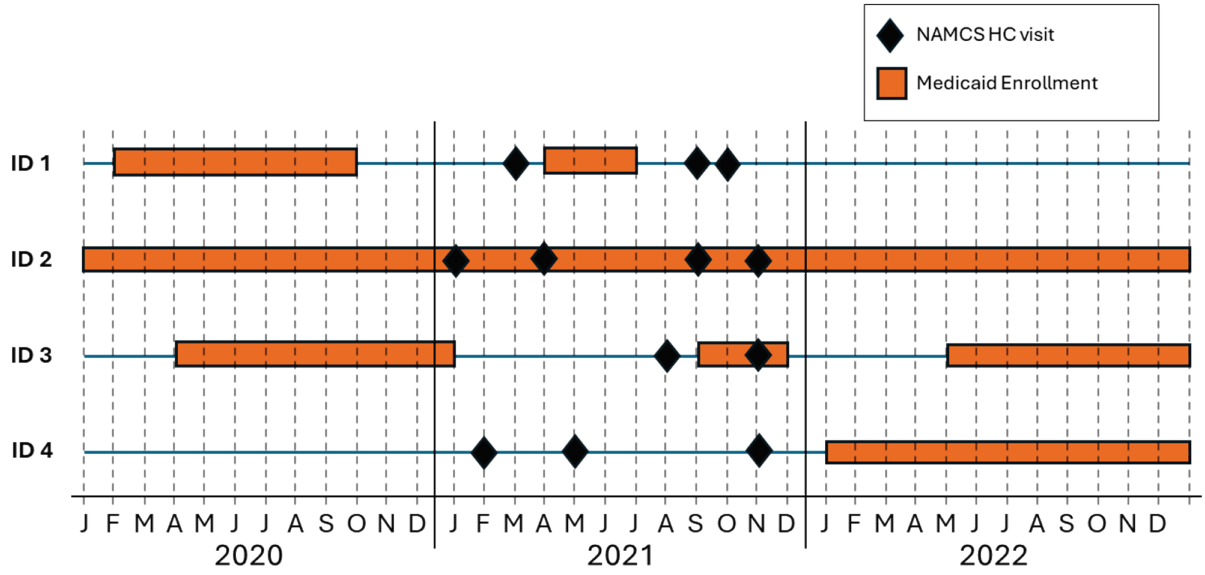
For more information on analyzing the 2021 NAMCS HC Component Visit Weights and monthly 2021 Medicaid enrollment days see Section 4.1.2.

Analysts proposing to analyze linked 2021 NAMCS HC Component- 2020-2022 T-MSIS data should request access to the Demographic and Eligibility (DE) TAF for the same calendar years as the TAF claims files (IP, LT, RX, OT) in order to determine the correct study denominators for the linked Medicaid population. The variable FILE_YEAR4 is used to identify the calendar year (2020-2022) of interest.

4.3.3 Temporal Alignment of Medicaid or CHIP Enrollment with 2021 NAMCS HC Component Patient Data

The linked data files include monthly Medicaid (MDCD_ENRLMT_DAYS_01 Thru 12) or CHIP enrollment (CHIP_ENRLMT_DAYS_01 THRU 12) variables for calendar years 2020 through 2022. Each variable provides a count of the number of days of Medicaid or CHIP enrollment during the 12 months of each calendar year. Therefore, Medicaid or CHIP enrollment status may be available for 2021 NAMCS HC Component patients during the calendar year prior to, the year of, or the year after the NAMCS HC Component data reporting period. Expanding on Figure 1, which shows four hypothetical patient visit and Medicaid or CHIP coverage scenarios for 2021, Figure 2 provides examples of how 2020-2022 Medicaid or CHIP coverage may align with 2021 NAMCS HC Component patient data during the 2020-2022 T-MSIS linkage period.

Figure 2. Temporal alignment of 2021 NAMCS HC Component patient visit data linked to 2020-2022 Medicaid enrollment data.



Sources: 2021 NAMCS HC Component patient data linked to 2020-2022 T-MSIS administrative data.
Note: T-MSIS is the Transformed Medicaid Statistical Information System.

The examples shown in Figure 2 are as follows,

- Patient ID1 was enrolled in Medicaid from February through September 2020, during the calendar year prior to the 2021 NAMCS HC Component data reporting period, as well as during the 2021 NAMCS HC Component data reporting period (April 2021 - June 2021); but was not enrolled in Medicaid during any month of the calendar year following the 2021 NAMCS HC Component data reporting period.
- Patient ID2 was enrolled in Medicaid throughout the entire 2021 NAMCS HC Component – 2020-2022 TMSIS linked data period (January 2020 - December 2022) including at the time of each their 2021 NAMCS HC Component patient visits.
- Patient ID3 was intermittently enrolled in Medicaid with periods of Medicaid coverage prior to (April 2020 - December 2020), during (September 2021 - November 2021), and after (May 2022-December 2022) the 2021 NAMCS HC Component data reporting period.
- Patient ID4 was not enrolled in Medicaid during any of their 2021 NAMCS HC Component reported patient visits but was enrolled in Medicaid for all months of the calendar year following the 2021 NAMCS HC Component data reporting period.

All five 2021 NAMCS HC Component linked T-MSIS TAF files contain 3 years (2020-2022) of T-MSIS data. The variable FILE_YEAR4 is used to identify the calendar year (2020-2022) of interest.

4.3.4 Multiple DE Records in the Same Calendar Year for Linked Survey Participants

Linked 2021 NAMCS HC Component – 2020-2022 T-MSIS patients may have multiple DE TAF records. Most often, this is because a patient is linked to more than one year of T-MSIS data. However, a patient may also be linked to multiple DE records within the same calendar year due to changes in program eligibility, moves between states, and less frequently, administrative data errors or linkage errors.

Multiple DE records within the same calendar year may have overlapping months of Medicaid enrollment data. For example, it is possible for a linked 2021 NAMCS HC Component patient to be enrolled in Medicaid in one state for part of the month, move to another state and enroll in Medicaid in the new state of residence during the same month. In considering how to assess Medicaid enrollment in the presence of multiple DE records within a year, analysts may wish to consider the use of various data elements that provide information about Medicaid eligibility and enrollment by month. Analysts should review the full list of monthly Medicaid eligibility and enrollment variables in the DE TAF codebook and request all variables of potential interest in their RDC application. More detailed analytic considerations regarding Medicaid eligibility and enrollment variables in the DE TAF is provided by CMS at https://resdac.org/sites/datadocumentation.resdac.org/files/2021-01/TAF_TechGuide_DE_File.pdf.

When there are multiple DE TAF records linked within the same calendar year, the variables PATIENT_ID, MSIS_SEQN, SUBMTG_STATE_CD, and FILE_YEAR4 must be used to link the health care claims records from the TAF claims files (IP, LT, RX, OT) with the appropriate DE TAF record. Analysts should request these variables in their RDC application for all TAF files they intend to analyze.

4.3.5 Identifying Special Populations

4.3.5.1 Pregnant Women

The 2021 NAMCS HC Component restricted-use visit data file includes a variable (PREGNANT) that identifies whether the health center visit was by a pregnant female regardless of reason for visit. CMS has provided technical specifications for "Identifying Pregnant and Postpartum Beneficiaries in Medicaid and CHIP Administrative Data"[12]. These specifications are based on a list of procedure, revenue and diagnosis codes from TAF files (DE, IP and OT) applied to female beneficiaries ages 8 to 64 years who were ever enrolled in Medicaid or CHIP during the calendar year.

4.3.5.2 Persons with Chronic Health, Mental Health, and Potentially Disabling Conditions

CMS has developed algorithms to assist analysts in identifying 41 chronic health, mental health, and potentially disabling conditions based on selected International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM), Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS) codes found in T-MSIS claims data[13]. For example, condition algorithms include coding specifications for identifying alcohol and substance abuse conditions including opioid use disorders and intellectual disabilities, learning disabilities, and autism spectrum disorder using TAF files. The 2021 NAMCS HC Component restricted-use condition data file includes a variable, CONDITION_CODE_R, that provides ICD-10-CM diagnosis codes extracted from patient visit records.

[12] https://www.medicaid.gov/medicaid/data-and-systems/downloads/macbis/mih_techspecs.pdf
[13] https://www2.ccwdata.org/web/guest/condition-categories-other.

5 Access to Data Files

5.1 Access to the Restricted-Use Linked 2021 NAMCS HC Component – 2020-2022 T-MSIS 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 accessible through the NCHS RDC Network for approved applications. Analysts who wish to access 2021 NAMCS HC Component restricted-use data files and the linked 2021 NAMCS HC Component-2020-2022 T-MSIS data files must complete an RDC application. The RDC staff will review all submitted applications to determine if the proposed project is feasible and to identify any potential disclosure risks. More information regarding the NCHS RDC Network and the RDC application process is available from: https://www.cdc.gov/rdc/.

5.2 Merging 2021 NAMCS HC Component Restricted-Use Data with Linked 2021 NAMCS HC Component-T-MSIS Data Files

The linkage between the 2021 NAMCS HC Component data and the 2020-2022 T-MSIS data was conducted at a patient level using patient level identifiers. The shared variable, PATIENT_ID, will be used by the RDC to merge 2021 NAMCS HC Component restricted-use data files with the 2021 NAMCS HC Component - 2020-2022 T-MSIS linked data files. Analysts should request the variable PATIENT_ID in their RDC application.

5.3 Requesting Linked 2021 NAMCS HC Component-2020-2022 T-MSIS Data Files and Variables

Analysts should request all variables of interest from both the 2021 NAMCS HC Component restricted-use data files and the 2021 NAMCS HC Component- 2020-2022 T-MSIS linked data files in their RDC application.

To obtain information on 2021 NAMCS HC Component patient eligibility for linkage and T-MSIS match status, analysts should request access to the variables ELIGSTAT and TMSIS_MATCH_STATUS from the 2021 NAMCS HC Component – 2020-222 T-MSIS Match Status file. (See Section 4.3.1 for more information regarding the variables available on the CMS T-MSIS Match Status file).

A complete set of T-MSIS TAF codebooks, providing information on the variables for each of the linked T-MSIS TAFs (DE, IP, LT, RX, OT), has been created to assist analysts in the variable selection process. Each of the TAF claims file (IP, LT, RX, OT) lists the variables from each of the claim file types (header, line, and occurrence. Analysts must specify both the specific TAF and the claim file type (header, line, or occurrence) for each requested variable in their RDC application.

Analysts proposing to analyze 2021 NAMCS HC Component - 2020-2022 T-MSIS linked data should request access to the relevant variables from the Demographic and Eligibility (DE) TAF for the same calendar years as the TAF claims files (IP, LT, RX, OT) to align health care claims data with the associated program enrollment variables by calendar year (variable FILE_YEAR4). In addition, analysts should request the variables PATIENT_ID, MSIS_SEQN, SUBMTG_STATE_CD, and FILE_YEAR4 from all linked TAF

files (DE, IP, LT, RX, OT) to associate health care claims records with the appropriate DE TAF demographic and enrollment record within each calendar year.

Analysts interested in examining Medicaid and CHIP program differences by state should request the variable SUBMTG_STATE_CD from each of the TAF files (DE, IP, LT, RX, OT) they intend to analyze.

All five 2021 NAMCS HC Component linked TAF files contain 3 years (2020-2022) of T-MSIS data. Analysts should request the variable FILE_YEAR4 in their RDC application and use this variable to identify the specific calendar year of T-MSIS data they wish to analyze.

5.4 Additional Related Data Sources

5.4.1 Linked 2021 NAMCS HC Component-NDI Mortality Files

Analysts interested in studying mortality among the 2021 NAMCS HC Component patient population may wish to incorporate 2021 NAMCS HC Component mortality data available in the linked 2021 NAMCS HC Component -2021-2022 NDI Mortality files. The linked mortality files include information on deaths identified for the entire 2021 NAMCS HC Component patient population through linkage with the National Death Index and are not limited to deaths among 2021 NAMCS HC Component patients who were linked to T-MSIS data. The linked mortality file includes Patient ID, date of birth, date of death, and cause of death information for linked 2021 NAMCS HC Component patients. Analysts should request all relevant linked mortality variables, including PATIENT_ID, in their RDC application.

5.4.2 Linked 2021 NAMCS HC Component- Housing and Urban Development (HUD) Administrative Data Files

Analysts interested in examining health outcomes related to housing assistance may also request variables from the linked 2021 NAMCS HC Component–2020-2022 HUD Administrative Data file. The linked HUD administrative data files include variables pertaining to the 2021 NAMCS HC Component patient's participation in HUD administered Housing Choice Voucher (HCV), Public Housing (PH), and/or Multifamily (MF) programs from 2020 through 2022. Analysts should request all relevant linked HUD data, including PATIENT_ID, in their RDC application.

Appendix I: Detailed Description of Linkage Methodology

1 2021 NAMCS HC Component and 2020-2022 T-MSIS 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 2021 NAMCS HC Component patient records and for T-MSIS demographic & enrollment (DE) 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 2021 NAMCS HC Component patient or T-MSIS beneficiary:

- SSN (validated)[14]
- DOB (month, day, and year)
- Sex
- 5-Digit ZIP code and state of residence
- First, middle initial, and last name

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

- Date values: when invalid or outside of expected range
- Sex values: when multiple sex values are recorded for the same person
- 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 2021 NAMCS HC Component patient and T-MSIS linkage submission files based on variation of the linkage variables. Similar to the cleaning process, a more elaborate routine was used to generate alternate records involving the name fields. Alternate records were generated for patients according to the following rules.

- Sex was missing. Two alternate records (one with male sex and the other with female) were created (note that this would result in having generated records run through both male and female specific linkage passes, and resulting duplicated links would be subsequently resolved.

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

- SSN with less than nine digits. A single alternate record was created where leading zeros were added to SSN values of length 7 or 8 to make a 9-digit SSN. Note, no alternate record was created if an invalid SSN would be created by adding 0's.
- Improbable date of birth. Age at time of survey was computed by subtracting the year of the survey and the year of birth. Records with age greater than 114 had a single alternate record created,
- If month and day were suspected of being imputed (ex. Jan 1st or June 15th), entire DOB was changed to missing[15]
- Otherwise, only year was changed to missing
- State of residence outside of U.S. and not in rest of world (RW) list. Alternate record was created with ZIP and state codes changed to missing
- ZIP code represents a different state. Using the ZIPSTATE() SAS function, state was imputed using the non-missing ZIP code. If the imputed state was different from the recorded state of residence, an alternate record using imputed state was created
- 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 2 below provides three examples of how alternate records were generated for nick names (Patient ID 1) and multiple part names (Patient ID 2 & 3), using hypothetical data. For patient 2, the first name was used to generate multiple records, and for patient 3, the last name was used.

Table 2. Example of Alternate Record Generation using Name Fields

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

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

Submission files, which combined the cleaned and validated PII fields, were created separately for 2021 NAMCS HC Component patient records and for T-MSIS DE records. During this process, multiple submission records were created for each patient/beneficiary to show all combinations of the recorded values for these fields. That is, if a patient/beneficiary had two states–of-residence recorded and three dates-of-birth recorded and each of the remaining fields had only one variant, then a total of six submission records would have been created for the patient/beneficiary (see Table 3 for example).

[15] Note, the date values are often recorded when the actual value is unknown.

Submission records that did not meet the eligibility requirements (see Section 3.1 Linkage Eligibility Determination) were removed from the submission file.

Table 3. Example of Alternate Records Caused by Different PII Values

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

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

PII – Personally Identifiable Information.

Additional post processing steps were taken after the initial 2021 NAMCS HC Component and T-MSIS linkage submission files were created. First, records from both the 2021 NAMCS HC Component and T-MSIS submission files were separated according to the sex value (male or female). The probabilistic linkage method assumes independence between the PII variables used to score the potential links. Records in the submission files were separated by sex to avoid violating this assumption, especially when first and/or last name and sex would be used as blocking and/or scoring variables.

2 Deterministic Linkage Using Unique Identifiers

The deterministic linkage, which was the next step in the linkage process, used only the 2021 NAMCS HC Component and T-MSIS 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. Further, records in the 2nd pass had to have a non-missing first or last name **AND** a non-missing date of birth part (month, day, or year) to be eligible for deterministic matching using the last 4 of SSN. After records had been linked using SSN, the algorithm validated the deterministic links by comparing first name, middle initial, last name, month of birth, day of birth, year of birth, ZIP code of residence, and state of residence. If the ratio of agreeing identifiers to non-missing identifiers was greater than 50% (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 first or last name in agreement to be deemed a deterministic match. Of note, 2021 NAMCS HC Component patients 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 set of records resulting from the deterministic match process 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 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)."[16] Intuitively developed rules can be used to define the blocking criteria, however, for this linkage, variable values in the data being linked were used to inform the development of a set of blocking passes that efficiently join the datasets together (i.e., multiple, overlapping blocking passes are run, each using a different blocking key). By using these data to create an efficient blocking scheme (or set of blocking passes), a high percentage of true positive links were retained while the number of false positive links were significantly reduced. A supervised machine learning algorithm used the 'truth source' (see Appendix I section 2) as the validation dataset and the 2021 NAMCS HC Component and T-MSIS 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".[17] [18]

The machine learning algorithm produced 6 blocking passes to be used in the blocking scheme. Table 4 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. Further, if only the ZIP code of residence was used as a blocking variable, then state of residence was excluded from the list of scoring variables as it is implied to agree on all records.

[16] 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.

[17] 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.

[18] 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.

Table 4. Blocking and Scoring Scheme Used to Identify and Score Potential Links

Key Number	Blocking Key	Scoring Key
1	First name, month of birth, day of birth, year of birth	Middle initial, last name, state of residence, ZIP code of residence
2	Last name, month of birth, day of birth, year of birth	First name, middle initial, state of residence, ZIP code of residence
3	Month of birth, day of birth, year of birth, state of residence	First name, middle initial, last name, ZIP code of residence
4	First name, state of residence, ZIP code of residence	Middle initial, last name, month of birth, day of birth, year of birth
5	Last name, first name, month of birth	Middle initial, day of birth, year of birth, state of residence, ZIP code of residence
6	Last name, year of birth, state of residence	First name, middle initial, month of birth, day of birth, ZIP code of residence

3.2 Score Pairs

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

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

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

- First Name or First Initial (when applicable)
- Middle Initial
- Last Name or Last Initial (when applicable)
- Year of Birth
- Month of Birth
- Day of Birth
- State of Residence
- ZIP Code (conditional on state agreement)

Except for first and last name, agreement status was set to 1 if the NAMCS and T-MSIS values for a particular PII variable agreed exactly, 0 if they disagreed, and missing (i.e., '.') if either value was missing on the paired records. The agreement status assignment for first and last name is explained further in section 3.2.2 of this appendix.

3.2.1 M and U Probabilities

The M-probability is the probability that the identifiers on a pair of records agree, given that records represent the same person (i.e., the records are a match). M-probabilities were estimated separately within each individual blocking pass and were calculated for each of the identifiers used for scoring (Table 4). 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 all digits matching (i.e., all 9 for records with 9-digit SSN or all 4 for 4-digit SSN). Further, to account for the alternate submission records generated during the creation of the submission files, the "best" agreement was taken for each of the scoring variables among the blocked records for each 2021 NAMCS HC Component patient ID and T-MSIS ID (see Tables 5 and 6 for example of alternate record summarization). Table 5 is an example of how the agreement flags for each of the scoring variables in Blocking pass 5 are created. A value of 1 means the information in the variable is exactly matching, while a 0 means they are not. Table 6 then represents how the multiple submission records in Table 5 are summarized into one record for each 2021 NAMCS HC Component patient and T-MSIS 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 6 are then used to estimate the M-probabilities for each of the specific scoring variables.

Table 5. Example of Agreement Flags Using Blocking Pass 5

Person Identifiers		PII Agreement flags[1]				
Patient ID	**T-MSIS ID**	**Middle Initial**	**Day of birth**	**Year of birth**	**ZIP Code**	**State of residence**
1	1	1	0	1	0	.
1	1	.	1	1	0	0
1	1	1	0	1	0	0
2	2	1	0	1	0	0
3	789	1	1	.	0	1
3	789	0	1	0	1	1
3	789	.	1	0	1	.
3	789	0	0	1	1	1
3	322	1	0	1	1	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 6. Example Showing Summarization of Blocked Record Pairs for M-Probability Estimation, based on Table 5 example

Person Identifiers		PII Agreement flags[1]				
Patient ID	T-MSIS ID	Middle Initial	Day of birth	Year of birth	ZIP Code	State of residence
1	1	1	1	1	0	0
2	2	1	0	1	0	0
3	789	1	1	1	1	1
3	322	1	0	1	1	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 and ZIP code 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
- ZIP Code of residence – because ZIP codes are dependent on the state in which they are located, only the records where state of residence agreed were used in the computation of the ZIP code M-probability (i.e., if state was not in agreement, then it would be assumed that ZIP code would also not agree)

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 day of birth in blocking pass 5, record pairs that did not agree on SSN that had a majority (i.e., greater than 50%) of the PII among middle initial, year of birth, state of residence, 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 2021 NAMCS HC Component and T-MSIS submission records agreed.

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

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

For example, if for blocking pass 1, the state of residence code for Florida ("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. Further, if there was no agreement in the 'catch-all' category, the U-probability would have been set to 0. To avoid a U-probability of 0, the 'catch-all' U-probability was computed by halving the minimum (i.e., lowest) U-probability among the individual value's U-probabilities. Further, if no individual value received a U-probability (i.e., all values assigned to 'catch-all') and there was no agreement, then the U-probability was set to 0.0001. For example, if the minimum U-probability among state of residence codes was 0.052 and there was no agreement among the catch-all records, the catch-all U-probability for state of residence would be 0.026 (0.052/2). If no state of residence code received a U-probability and there was no agreement, the U-probability for state of residence code would be 0.0001. The process for calculating U-probabilities for first and last name differs from these methods and is described in Section 3.2.2.

Lastly, an adjustment was made to the final U-probabilities to account for alternate records in the submission file. With the addition of each alternate record, the chance of agreement between the 2021 NAMCS HC Component and T-MSIS submission records increases. For example, a 2021 NAMCS HC Component patient with different months of birth reported on two different patient visit records, has twice the chance of linking to a T-MSIS submission record. Therefore, the U-probability for that patient's month of birth should represent the combined chance of agreement across both month values. Section 3.2.3 provides a detailed description of the methods used to adjust the U-probabilities to account for the additional alternate submission records.

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 2021 NAMCS HC Component record was "Albert" and on the T-MSIS record it was "Abert", this comparison would receive a Jaro-Winkler score of 0.96. M-probabilities are computed as the rate of agreement for all first/last names within a specific Jaro-Winkler level. For example, the M-probability for first name at the Jaro-Winkler 0.90 level is the rate of agreement for all first names with a Jaro-Winkler score of 0.90 and above.

Because of the large number of unique name values, it was impractical to compute name specific U-probabilities for each blocking pass (i.e., there would not be enough records available for it to be done accurately). Instead, U-probabilities were estimated using pairs generated by the Cartesian product of all records in the 2021 NAMCS HC Component linkage submission file and 100,000 randomly selected names from a simple random sample of 20% of records with non-missing name information from the T-MSIS submission file. See Table 7 for the number of sampled T-MSIS submission records.

Table 7. Count of Records from a 20% Simple Random Sample of T-MSIS Submission Records used to Estimate U-Probabilities for First and Last Names by Sex

	Count of Sampled Records by Name	
Sex	**First Name**	**Last Name**
Female	29,517,200	29,782,942
Male	23,230,809	23,486,860

Complete name tallies (separately, for first and last names) were then produced for the 2021 NAMCS HC Component linkage submission file. 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 T-MSIS file names were then tallied.[19] [20] [21]

3.2.3 Adjustment of U-Probabilities for Alternate Submission Records

As previously mentioned in section 3.2.2, an adjustment was made to the U-probabilities to account for alternate submission records. The addition of unique values for an identifier increases the likelihood of a spurious linkage between records from the files being linked. Thus, the U-probabilities were adjusted to account for the increased probability of variable agreement (i.e., if records for the same person had multiple values for a variable, the chance of agreement with any compared record from the other file increases). Therefore, patients received an adjusted U-probability if they had identifier values that were different across their set of submission records. The adjusted U-probabilities were then applied to each record in the set of submission records that paired with a T-MSIS DE record. Lastly, the U-probability that is used to compute the agreement and disagreement weights (see Section 3.2.4) is the maximum between the original and adjusted U-probability (i.e., $U_{Max}=Max(U_{Original}, U_{Adjust})$).

Excluding first and last name and ZIP code of residence, the adjustment process began by identifying the unique set of values, and their U-probabilities, for each of the identifiers appearing in the scoring key (Table 4), for each patient. Because each value is assumed to be independent of the others, the adjusted U-probabilities were computed using the additive rule for probability as the summation of the individual value U-probabilities for each patient. That is, if a patient had three different month of birth values, the adjusted U-probability for month of birth was simply the summation of the three individual U-probabilities. Table 8 provides an example of the process used to compute the adjusted and maximum U-probabilities for month of birth.

[19] 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.
[20] 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.
[21] 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.

Table 8. Example Showing Computation of the Adjusted and Maximum U-probability for Month of Birth

Patient ID	Month of Birth	U-Probability	Adjusted U-Probability[1]	Maximum U-Probability[2]
1	6	0.091		0.253
1	5	0.083	0.253	0.253
1	7	0.079		0.253
2	1	0.110	0.191	0.191
2	10	0.081		0.191
3	6	0.091	0.091	0.091

NOTES: Data have been fabricated for the purposes of this example

[1] The adjusted U-probability is computed by summing the individual month of birth U-probabilities by patient ID.

[2] The maximum U-probability is the max U-probability value between the original and adjusted U-probabilities.

The first three columns of Table 8 show the unique values of month of birth and their corresponding U-probabilities (see Section 3.2.1) for patients 1, 2, and 3. The column titled "Adjusted U-Probability" is computed by totaling the individual probabilities in the third column for each patient. Finally, the maximum U-probability (last column), which was used to compute the agreement and disagreement weights (see Section 3.2.4), is the maximum value between the original and adjusted U-probability values.

Because ZIP codes are nested within the state of residence codes, a slightly different process was used to compute the adjusted U-probability for ZIP code. The process began by identifying the unique set of state and ZIP of residence codes, along with the U-probability for each ZIP code, for each patient. Next, each of the U-probabilities for ZIP code of residence were summed to the patient and state of residence level. Finally, the patients adjusted U-probability for ZIP code was computed as the average of the summed U-probabilities for ZIP codes across the reported state of residence codes. The computation of the adjusted U-probability for ZIP code of residence can be represented by the following equation,

$$U_{Adjust\ ZIP} = \frac{\sum_{i=1}^{n}(\sum_{j=1}^{m} U_j)}{n}$$

where n is the number of unique state codes, m is the number of unique ZIP codes, and U_j is the U-probability for the j^{th} ZIP code. Table 9 provides an example of the process used to compute the adjusted U-probability for ZIP code of residence.

Table 9. Example Showing Computation of the Adjusted and Maximum U-probability for ZIP Code of Residence

Patient ID	State of Residence	ZIP Code of Residence	U-Probability	Adjusted U-Probability[1]	Maximum U-Probability[2]
8	CA	90002	0.001		0.0047
8	CA	90313	0.003	0.0047	0.0047
8	FL	32011	0.01		0.01
25	GA	31013	0.001	0.0015	0.0015
25	GA	39845	0.002		0.002
78	CT	06752	0.001	0.001	0.001

NOTES: Data have been fabricated for the purposes of this example

[1] The adjusted U-probability is computed by summing the individual ZIP code U-probabilities within each state code and then taking the average of the summed U-probabilities across the states for each patient ID.

[2] The maximum U-probability is the max U-probability value between the original and adjusted U-probabilities. Recall, the maximum U-probability is the maximum U-probability value between the original (column 4) and adjusted (column 5) U-probabilities.

The first four columns of Table 9 provide the Patient ID, state of residence, ZIP code of residence codes, and the corresponding U-probability for each ZIP code of residence for three fabricated 2021 NAMCS HC Component patients. The adjusted U-probability (i.e., 5th column) is computed first by summing each individual U-probability within each state code and then taking the average of the summed values. The maximum U-probability (i.e., last column) is the max U-probability value between the original and adjusted ZIP-code of residence U-probabilities. Notice, for patients 8 and 25, the maximum U-probability value that was used for ZIP code 32011 and 39845, respectively, was the original U-probability. This was because the average U-probability across all state codes (column 5) did not exceed the original U-probability (column 4).

For first and last names, only the 85% Jaro-Winkler level U-probability was adjusted. The higher levels (i.e., 90, 95, and 100) were not adjusted because of the hierarchical method being used to compute each of the U-probabilities at those levels (i.e., 90 is dependent on 85, 95 is dependent on 90, and 100 is dependent on 95). Before the 85% level was adjusted, names that were similar to one another were combined into a single name field. This step is necessary to avoid 'double counting' names that are highly likely to match to the same name on the T-MSIS DE data file. Similarity in names was defined as having a Jaro-Winkler score between 0.95 and 1 (not inclusive at the upper bound) or if one name is fully contained within another (ex. Elizabeth and Eliza). If for example, a patient had two different names, Elizabeth and Elizabith (JW_{Score}=0.967), only one would be used to adjust the 85% Jaro-Winkler U-probability. The name that is selected was determined by whichever had the highest 100% Jaro-Winkler U-probability. Using the list of 'unduplicated' names, the adjusted U-probability for the 85% Jaro-Winkler level was computed as the summation of each of the individual U-probabilities for the patient. Table 10 provides an example of the methods used to compute the adjusted U-probabilities for the 85% Jaro-Winkler level, using first name as an example.

Table 10. Example Showing Computation of the Adjusted and Maximum U-probability for First Name

Patient ID	First Name	U-Probability at 85% JW	U-Probability at 100% JW	Collapsed U-Probability[1]	Adjusted U-Probability[2]	Maximum U-Probability[3]
8	Margaret	0.008	0.99	0.008		0.009
8	Peggy	0.001	0.97	0.001	0.009	0.009
8	Marg	0.001	0.85	Collapsed		0.009
25	Elizabeth	0.09	0.99	0.09	0.09	0.09
25	Beth	0.01	0.95	Collapsed		0.09
78	Cathy	0.05	0.99	0.05	0.05	0.05

NOTES: Data have been fabricated for the purposes of this example. JW is the Jaro-Winkler string comparator function.
[1] The collapsed U-probability includes only the U-probabilities after similar names have been collapsed into a single name.
[2] The adjusted U-probability is computed by summing each of the collapsed 85% JW U-probabilities within each patient ID.
[3] The Maximum U-probability is the max U-probability value between the original and adjusted 85% U-probabilities.

The first four columns of Table 10 provide example Patient IDs, first names, and their U-Probabilities at the Jaro-Winkler 85 and 100 level for three fabricated 2021 NAMCS HC Component patients. The collapsed U-probability column (i.e., 5th column) shows that two names were collapsed into another, i.e., for patient 8, Marg was collapsed into Margaret (full-containment) and Beth was collapsed into Elizabeth (full-containment) for patient 25. Further, the collapsed U-probability is equal to the 85% JW U-probability for the name with the highest 100% JW U-probability among the names being collapsed. The adjusted U-probability (i.e., column 6) is the summation of each collapsed U-probability for each patient ID. Finally, the maximum U-probability (i.e., last column) is the max value between the adjusted U-probability and original U-probability at the 85% JW level.

3.2.4 Calculate Agreement and Non-Agreement Weights

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

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

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

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

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

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

3.2.5 Calculate Pair Weight Scores

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

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

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

3.3 Probability Modeling

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

- Combine pairs across blocking passes (Pair-weights are specific to each blocking pass and are not comparable)
- Select a "best" record among 2021 NAMCS HC Component patient IDs that have linked to multiple administrative records.
- Select final matches based on a probability cut-off value (discussed in the following Section 4)

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

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

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

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

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

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

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

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

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

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

Continuing with the example…
For Pair 1 in blocking pass B, $P_{EM,P,B}(Match) = \left(\frac{6.9}{6.9+1}\right) \approx 0.87$
For Pair 2 in blocking pass $P_{EM,P,B}(Match) = (\frac{0.0036}{0.0036+1}) \approx 0.0036$
…and this continues for the remaining $N_{Pairs,B}$ pairs of the blocking pass.

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

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

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

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

This process was repeated until convergence was reached in the number of matches being estimated. Once convergence was achieved, the final probabilities were estimated based on the last value of $\widehat{N_{matches,B}}$ to be estimated. These estimated probabilities were then used to select the final matches, as described below in Section 4.

3.4 Adjustment for SSN Agreement

Up to this point, every pair generated through the probabilistic routine was assigned a value that estimates its probability of being a match. However, this estimate did not take SSN agreement into account. This was conducted as a separate step because for the other comparison variables, M- and U-probabilities were estimated based on probable matches or non-matches that were determined based on SSN agreement, and clearly this was infeasible for SSN itself.[22]

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 2021 NAMCS HC Component and T-MSIS submission record, the estimated probability was adjusted based on the last four digits of the SSN.

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

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

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 2021 NAMCS HC Component patient or T-MSIS submission record. So, for these pairs:

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

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

The scored (probabilistic) and deterministic linkage files for males and females were combined prior to estimating the linkage error and selecting matches. Recall the purpose for separating the records by sex was to avoid violating the independence assumption for name identifiers mentioned by Fellegi-Sunter. Now that records from each sex have been separately scored, there is no need to keep them separate.

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

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 the last four digits matching, regardless of SSN length) 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 2021 NAMCS HC Component and T-MSIS 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%.

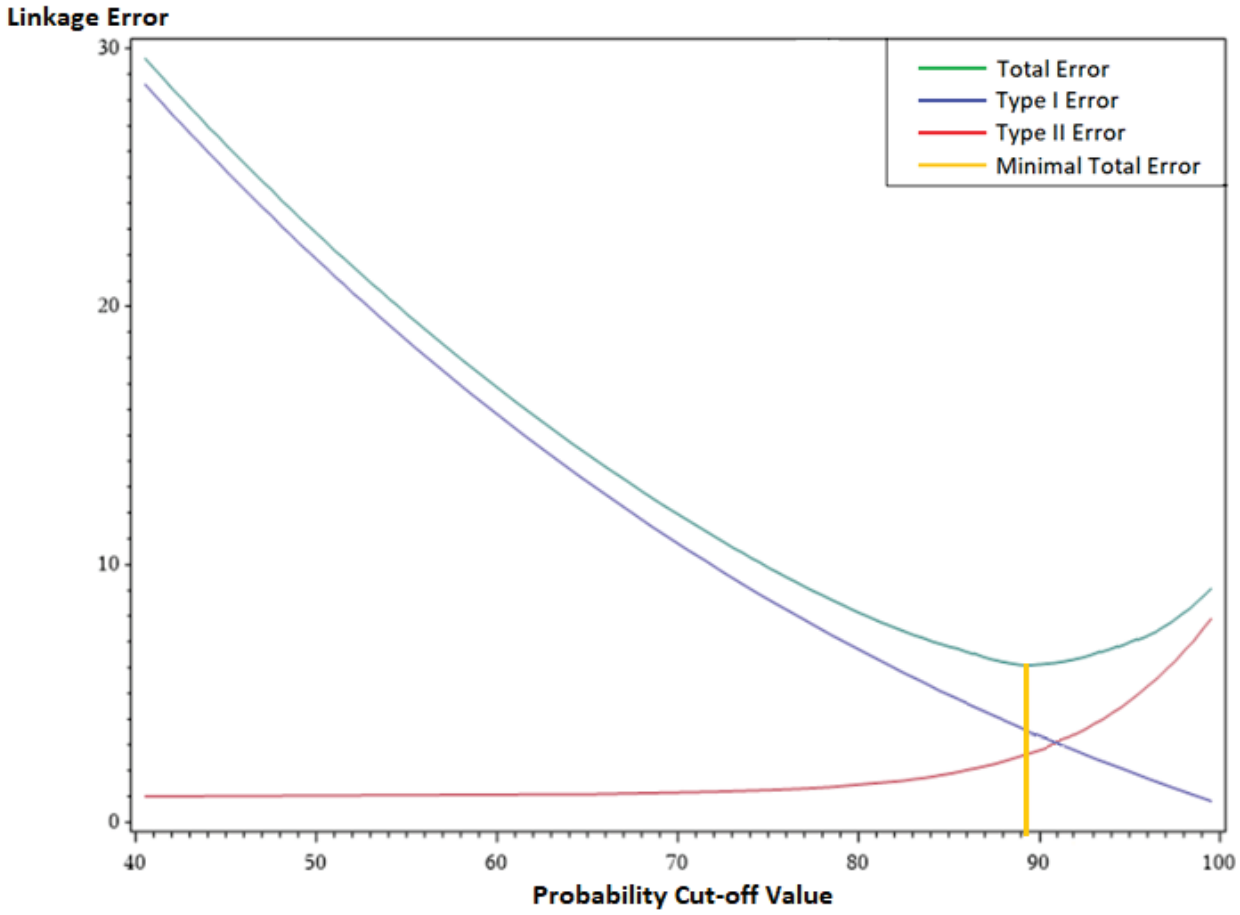
To measure Type II error, the truth source comprised of all matches identified in the deterministic linkage was used. Recall, the truth source contains records with full nine-digit SSN agreement (step 1) or with the last four digits of SSN in agreement (step 2). Potential deterministic matches were then 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 is selected (see Figure 3). For the linkage of the 2021 NAMCS and 2020-2022 T-MSIS, the optimal probability cut-off value was set to 0.85.

Figure 3. 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). 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 11 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 2021 NAMCS HC Component T-MSIS 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 2021 NAMCS HC Component record was a match to the T-MSIS DE 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 11. 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)
2021 NAMCS HC Component	0.85	414,862	215,859 (52%)	199,003 (48%)	<0.1%	0.3%

Appendix II: Descriptions of TAF Files

Additional information about the TAFs is available from CMS. Analysts should also consult www.ccwdata.org and www.resdac.org for more specific information about how to analyze TAF data including accessing the "T-MSIS Analytic Files (TAF) Research Identifiable Files (RIFs) User Guide" available via www.ccwdata.org.

1 Demographic and Eligibility (DE) File

This file provides demographic and program eligibility and enrollment information on each person who was enrolled for at least one day in the calendar year in Medicaid and/or CHIP. Demographic data elements in the DE file include race and ethnicity, primary language, and marital status. Eligibility data elements include Medicaid and CHIP enrollment days, eligibility group, CHIP program (either M-CHIP or S-CHIP), dually eligible individual status, restricted benefit status, and participation in managed care, for each month in the calendar year. The file also includes information about the enrollee's participation in other federal programs, such as Social Security Disability Insurance (SSDI), Supplemental Security Income (SSI), and Temporary Assistance for Needy Families (TANF).

2 Inpatient Hospital (IP) File

This file includes records for inpatient hospital services for Medicaid and CHIP enrollees during the calendar year. Emergency room visits that result in an inpatient hospital admission are identified in Uniform Billing (UB-04) revenue codes (T-MSIS claim line-item data element REV_CNTR_CD). Prescribed drugs, supplies and other items provided by a hospital's pharmacy are aggregated in UB-04 codes. Emergency room visits that do not result in an inpatient hospital admission are not included in this file but are reported in the Other Services (OT) file.

3 Long-Term Care (LT) File

This file includes records for institutional long-term care services for Medicaid and CHIP enrollees during the calendar year. Records include claims for room and board, which may include prescribed drugs if they are included in the institution's per diem rate. LT records also include ancillary services, such as speech therapy or specialized dietary services, if they are provided by the institution's staff. Otherwise, prescribed drugs and ancillary services are reported in the RX and OT files, respectively.

4 Pharmacy (RX) File

This file includes records for prescribed drugs, supplies, and other items provided by a free-standing pharmacy, either directly to an enrollee or to a long-term facility for the enrollee's use. This includes prescribed and covered over-the-counter drugs, supplies, and durable medical equipment. Injectable drugs (such as immunizations) administered by a health professional in a physician's office, group practice, or clinic are reported in the OT file. Records for immunizations provided at free-standing pharmacies are included in the RX file. Note, it is possible for RX header claims to have no corresponding record in the RX line file. When sufficient information exists on the RX header claim to describe the drug prescription/dispensing, no line record is required.

Medicaid payment amounts for prescribed drugs are reported prior to the receipt of manufacturer rebates. Pharmacy records include National Drug Codes (NDC), but for analysis of prescription drug use, an NDC[23] does not identify the primary therapeutic use of a drug. Analysts who need to determine the primary therapeutic use of a given NDC will need to link NDCs from the RX file (T-MSIS data element NDC) to external sources of information. Analysts should identify any external sources of information to be used in their analyses in their RDC application (see Section 5.1 for additional information).

5 Other Services (OT) File

This file includes records for all other community-based services not reported in the IP, LT, and RX files. These services include physicians (including separately billed services provided to patients during inpatient hospital stays); clinic; laboratory; radiology; Early and Periodic Screening, Diagnostic and Treatment Services; home health; dental; therapy; transportation; case management; family planning services; waiver services; and home and community-based services. As noted above, this file includes records for emergency room services that do not result in a hospital admission, some immunizations, and injectable drugs that must be administered by a medical professional, except as noted above. This file also includes monthly premium payments made by the state Medicaid program to prepaid managed care plans.

[23] https://www.fda.gov/drugs/drug-approvals-and-databases/ndc-product-file-definitions