The Linkage of the 2016 National Hospital Care Survey to 2015–2017 Centers for Medicare & Medicaid Services Transformed Medicaid Statistical Information System Claims Data: Matching Methodology and Analytic Considerations

Data Release Date: July 15, 2022

Document Version Date: July 15, 2022

Division of Analysis and Epidemiology

National Center for Health Statistics

Centers for Disease Control and Prevention

datalinkage@cdc.gov

Suggested Citation: National Center for Health Statistics. Division of Analysis and Epidemiology. *The Linkage of the 2016 National Hospital Care Survey to 2015–2017 Centers for Medicare & Medicaid Services Transformed Medicaid Statistical Information System Claims Data: Matching Methodology and Analytic Considerations*, July 2022 Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data-linkage/index.htm

Contents

List of Acronyms

AIDS, Acquired Immunodeficiency Syndrome

CFR, Code of Federal Regulations

CHIP, Children’s Health Insurance Program

CMS, Centers for Medicare & Medicaid Services

DE, Demographic and Eligibility

DOB, date of birth

DQ, data quality

DRG, Diagnosis Related Group

DSH, Disproportionate Share Hospital

ED, emergency department

EHR, electronic health record

EM, expectation-maximization

EPSDT, Early and Periodic Screening, Diagnosis, and Treatment

ERB, Ethics Review Board

FFS, fee-for-service

FMAP, Federal Medical Assistance Percentage

FPL, Federal Poverty Level

FQHC, Federally Qualified Health Center

HCBS, Home- and Community-Based Services

HIO, Health Insuring Organization

HIV, Human Immunodeficiency Virus

HMO, Health Maintenance Organization

HUD, U.S Department of Housing and Urban Development

IMD, Institution for Mental Disease

IP, inpatient services

LT, long-term care services

MAX, Medicaid Analytic eXtract

MBSF, Master Beneficiary Summary File

MCO, managed care organization

MSIS, Medicaid Statistical Information System

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

NCHS, National Center for Health Statistics

NDC, National Drug Code

NDI, National Death Index

NHCS, National Hospital Care Survey

OP, outpatient

OPD, outpatient department

OT, Other services

PACE, Program for All-Inclusive Care for the Elderly

PAHP, Prepaid Ambulatory Health Plan

PCCM, Primary Care Case Management

PH, Public Housing

PHP, Prepaid Health Plan

PIHP, Prepaid Inpatient Health Plan

PII, personally identifiable information

PW, pair weight

PYE, Person year equivalent

QDWI, Qualified Disabled Working Individual

QI, Qualifying Individual

QMB, Qualified Medicare Beneficiary

RDC, Research Data Center

ResDAC, Research Data Assistance Center

RX, Prescription drug services

S-CHIP, State Children’s Health Insurance Program

SHADAC, State Health Access Data Assistance Center

SLMB, Specified Low-income Medicare Beneficiary

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

TOS, Type of Service

TPI, Transformed Medicaid Statistical Information System Priority Item

UB-04, uniform billing form

UPL, Upper Payment Limit

1 Introduction

As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to provide statistical information that can be used to guide actions and policy to improve the health of the American people. In addition to collecting and disseminating the Nation's official vital statistics, NCHS conducts several population-based surveys and healthcare establishment surveys, including the National Hospital Care Survey (NHCS), https://www.cdc.gov/nchs/nhcs/index.htm (accessed June 10, 2022). The NHCS collects electronic health records or health care claims data from participating hospitals drawn from a national sample frame of non-institutional and non-federal hospitals with six or more staffed inpatient beds. Participating hospitals are requested to send all patient ambulatory care and inpatient (IP) encounters occurring within the data collection calendar year. The NHCS includes detailed information about each participating hospital's patients' characteristics, conditions, and treatment. Even though NHCS is an establishment survey (i.e., hospitals are the sampling unit) it collects patient personally identifiable information (PII), which enable data linkages.

Through its Data Linkage Program, NCHS has been able to expand the analytic utility of the data collected from NHCS by augmenting it with Medicaid and Children's Health Insurance Program (CHIP) claims data collected by the Centers for Medicare & Medicaid Services (CMS) Transformed Medicaid Statistical Information System (T-MSIS). **This report will describe the linkage of data from the 2016 NHCS to 2015-2017 CMS T-MSIS claims data.** Linking data from the 2016 NHCS to 2015-2017 CMS T-MSIS data provides researchers with linked T-MSIS information one year prior, the year of, and one year after the NHCS data collection period. Although the 2016 NHCS data are not nationally representative due to low survey response rates, linking NHCS data with CMS T-MSIS claims creates a new data resource that can support research studies focused on a wide range of patient health outcomes and the association of means-tested government insurance programs on health and health outcomes.

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

The data linkage work was performed at NCHS in part through contract #HHSD2002016F92236B by NORC at the University of Chicago with funding from the Department of Health and Human Services' Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF).

2 Background on Linked Files

2.1 National Hospital Care Survey

The NHCS is an establishment survey that collects IP, emergency department (ED), and outpatient department (OPD) episode-level data from sampled hospitals. NHCS is one of the NCHS National Healthcare Surveys, a family of surveys that are provider-based, covering a broad spectrum of health care settings (https://www.cdc.gov/nchs/dhcs/index.htm). The goal of NHCS is to provide reliable and timely healthcare utilization data for hospital-based settings, including prevalence of conditions, health status of patients, and health services utilization.

NHCS collects data from participating hospitals on all IP and ambulatory care visits occurring during the calendar year for patients of all ages (including newborns). During the 2016 data collection, hospitals were given the option of providing their data in the form of electronic health records (EHRs) or as Uniform Bill (UB)-04 administrative claims records. Thus, participating hospitals provided data in the form of UB-04 claim records or EHR data, where the EHR data in 2016 were provided in the form of Continuity of Care Documents (CCDs) or custom extracts.

NHCS collects patient PII (e.g., full name, date of birth, and Social Security Number (SSN)), which allows for the linkage of episodes of care across hospital units as well as to other data sources, such as CMS T-MSIS data. The linkage described throughout this document only includes the linkage to T-MSIS data for patients with either IP or ED visits reported in NHCS – patients who only had other, non-ED OPD visits reported in NHCS have been excluded from the linkage.

The NHCS sample frame includes 6,622 non-institutional, non-federal hospitals with six or more staffed inpatient beds. A base sample of 500 hospitals and a reserve sample of 500 additional hospitals was drawn from this frame. In 2013, to provide estimates for ED visits with incidents of substance abuse, 81 hospitals with 500 or more staffed inpatient beds were added to the NHCS sample from the reserve sample. Thus, the hospital sample size for the 2016 NHCS data collection was 581 hospitals. In 2016, 158 out of the 581 sampled hospitals provided data and of the 158 participating hospitals, 142 were determined to be in-scope for linkage. Hospitals were determined to be out-of-scope for linkage if they did not provide patient PII, provided less than 50 patient encounter records or did not provide patient records covering at least 6 months of the data collection period. Of those 142 linkage eligible hospitals, 140 hospitals submitted IP data and 121 hospitals submitted ED data.

2.2 Centers for Medicare & Medicaid Services (CMS) Transformed Medicaid Statistical Information System (T-MSIS) Claims Data

2.2.1 Medicaid

Enacted in 1965 as Title XIX of the Social Security Act, Medicaid is a federal and state partnership to provide health insurance coverage to low-income individuals in the United States. The program has changed continuously since it was enacted through a series of legislative actions [1]. Medicaid is jointly financed with federal and state/local funds [1]. States must meet federal requirements to receive federal funding. Management and oversight activities are shared by federal and state governments, with identified federal and state roles and responsibilities. There is significant variation among state Medicaid programs in both the eligible population groups and the covered services. For this reason, each state must develop and maintain a state Medicaid plan to assure that the state abides by federal requirements for administering its program and claiming federal matching funds[2].

Over 85.8 million individuals were enrolled in Medicaid and CHIP in the District of Columbia and the 50 states that reported enrollment data for November 2021. Among that total, over 78.9 million individuals were enrolled in Medicaid and nearly 6.9 million individuals were enrolled in CHIP [2]. Enrollment in these programs represented over one quarter of the U.S. population[3], with over 90 million individuals enrolled for at least one day in 2018. Medicaid provided coverage for 42.3% of U.S. births in 2018 [3].

Medicaid accounts for almost one-sixth of national spending on personal health care [4]. Medicaid is the main payer of nursing home care and long-term care services overall [5]; it is also the largest source of public funding for mental health care [6]. Seniors and people with disabilities make up approximately 25% of all Medicaid enrollees but account for two-thirds of Medicaid benefit expenditures [7]. The Federal Medical Assistance Percentage (FMAP), also called the federal match rate, represents the percentage of Medicaid service expenditures financed by the federal government in each state. FMAP differs by state and is based on the average per capita income in a state relative to the national average. The combined federal and state/local shares of Medicaid spending were $688.0 billion in fiscal year 2020 [8], more than double the spending of $333.2 billion in fiscal year 2007. In fiscal year 2020, the federal share of spending was 67.0 percent while state and local spending accounted for the remaining 33.0 percent. Spending is estimated to exceed $1 trillion annually before 2030 [9]. Therefore, federal and state policy makers have been implementing strategies to contain spending growth while improving access, equity, and quality for program enrollees.

Medicaid is a means-tested health insurance program that provides health care coverage to certain mandated low-income populations [10], such as:

- Poverty-related eligibility for pregnant women and deemed newborns, infants, and children to age 18[4]
- Low-income families (with income below the state’s 1996 Aid to Families with Dependent Children limit, often below 50% of the Federal Poverty level (FPL))
- Families receiving transitional medical assistance
- Children with Title IV-E adoption assistance
- Foster care, or guardianship care, and individuals aging out of foster care [11]
- Elderly and disabled individuals receiving Supplemental Security Income (SSI)

[1] For a legislative history of Medicaid and CHIP, see https://www.macpac.gov/reference-materials/federal-legislative-milestones-in-medicaid-and-chip/.

[2] For more information on Medicaid state plan requirements, see https://www.macpac.gov/subtopic/state-plan/ (accessed June 10, 2022).

[3] The U.S. population estimate for July 1, 2018 was 327.2 million individuals, 2018 National and State Population Estimates (census.gov).

[4] Income cut-offs are based on percentage of the federal poverty level (FPL) and vary by state.

- Aged, blind, and disabled (any age) individuals in Section 209(b) states[5]
- Certain working individuals with disabilities
- Certain low-income Medicare enrollees (known as dually eligible individuals)
- Refugees and Asylees (including Afghan refugees)
- Undocumented immigrants (emergency services only)[6]

Optional eligibility groups[7] include the following low-income individuals:

- Originally, the 2010 Patient Protection and Affordable Care Act required all states to provide Medicaid coverage for low-income adults ages 21 to 64 (below 138% FPL), but this requirement was overturned by the Supreme Court in 2012. As of 2022, 38 states and the District of Columbia provide this coverage.
- As of 2022, 35 States and the District of Columbia provide Medicaid coverage under medically needy and medically needy spenddown provisions. Individuals qualify for medically needy provision coverage if their income falls below a state-imposed income threshold. Individuals can also qualify for medically needy spenddown provision coverage if their income minus medical costs falls below the state determined threshold.
- Women with breast and cervical cancer
- Certain individuals ages 19-20, including those residing in foster homes, those with subsidized adoptions, or those with intellectual disabilities who reside in intermediate care facilities, nursing homes, or psychiatric institutions.

As of November 2021, 85,809,179 individuals were enrolled in Medicaid and CHIP in the District of Columbia and the 50 states that reported enrollment data.

- 78,910,300 individuals were enrolled in Medicaid.
- 6,898,879 individuals were enrolled in CHIP.

To receive federal funding for Medicaid, states must offer enrollees a core set of mandatory services, although states can place limits on the amount, duration, or scope of services that enrollees receive. The mandatory services include:

- Inpatient hospital
- Nursing facility (age over 21) - no distinction between skilled and intermediate care levels
- Home health
- Outpatient hospital
- Rural health clinics

[5] Under Section 209(b), states may choose criteria other than receipt of SSI cash payments as a basis for granting Medicaid eligibility to low-income aged and disabled individuals. As of 2022, Connecticut, Hawaii, Illinois, Minnesota, Missouri, New Hampshire, North Dakota, Ohio, Oklahoma, and Virginia have chosen this option.

[6] Federal law generally bars undocumented immigrants from being covered by Medicaid, but the federal government has provided funds to states to cover emergency services for people who, other than their citizenship status, meet all other criteria for Medicaid eligibility through a program called Emergency Services for Aliens.

[7] For more information on eligibility, including mandatory and optional eligibility groups, see https://www.medicaid.gov/sites/default/files/2019-12/list-of-eligibility-groups.pdf (accessed June 10, 2022)

- Federally Qualified Health Centers (FQHCs)
- Physician
- Laboratory and x-ray
- Subject to state law or regulation:
    - Nurse midwife services
    - Certified pediatric or family nurse practitioner
- Freestanding birth centers
- Early and Periodic Screening, Diagnosis and Treatment (EPSDT)
- Family planning services and supplies
- Non-emergency medical transportation
- Tobacco cessation counseling

Federal matching funds are also available for any services identified on a list of optional services that states may choose to cover[8].

- Prescription drugs
- Clinic services
- Physical therapy
- Occupational therapy
- Speech, hearing, and language disorder services
- Respiratory care services
- Other diagnostic, screening, preventive, and rehabilitative services
- Podiatry services
- Optometry services
- Dental Services
- Dentures
- Prosthetics
- Eyeglasses
- Chiropractic services
- Other practitioner services
- Private duty nursing services
- Personal care
- Hospice
- Case management
- Services for individuals aged 65 or older in an Institution for Mental Disease (IMD)
- Services in an intermediate care facility for Individuals with Intellectual Disability
- State Plan Home and Community Based Services
- Self-Directed Personal Assistance Services
- Community First Choice Option
- Tuberculosis-related services
- Inpatient psychiatric services for individuals under age 21

[8] For complete lists of mandatory and optional services, see https://www.medicaid.gov/medicaid/benefits/index.html (accessed June 10, 2022).

- Health homes for enrollees with chronic conditions
- Other services approved by the Secretary

States may make changes to optional eligibility and service coverage provisions at any time during the calendar year.

2.2.1.1 Early and Periodic Screening, Diagnostic and Treatment (EPSDT)

An important feature of Medicaid is the EPSDT program for enrolled children [12]. The following EPSDT services are required of all Medicaid programs for children under the age of 21:

- Periodic health screenings
- All services necessary to correct or ameliorate physical or mental health conditions identified by a screening
- Vision services, including eyeglasses
- Dental services, including dental care, treatment to relieve pain and infections, restore teeth, and maintain dental health
- Hearing services, including hearing aids
- Any other medically necessary services listed in the Medicaid statute, including optional services that are not otherwise covered by the state

There is variation in the reporting of EPSDT services across the states. Some states report only EPSDT screenings and services provided via direct referrals from those screenings as EPSDT. Other states report nearly all services provided to enrolled children as EPSDT.

2.2.1.2 Requirements and Waivers

Medicaid programs must assure the following:

- Comparability – A Medicaid covered benefit generally must be provided in the same amount, duration, and scope to all enrollees.
- Freedom of choice – All enrollees must be permitted to choose a health care provider from among any of those participating in Medicaid.
- Statewideness – A Medicaid program cannot exclude enrollees or providers because of where they live or work in a state.

However, any or all three of these requirements can be waived if a state applies for a waiver and CMS approves the waiver [13]. The purpose of waivers is to allow exemptions to the requirements of comparability, statewideness, and freedom of choice. In general, waivers allow states flexibility to identify a specific set of services not otherwise required or optional for states, deliver services to a defined sub-population of Medicaid enrollees, target a substate area, mandate enrollment in managed care, and/or implement program innovations, such as alternative delivery systems. States must submit a waiver application to CMS and receive approval before they can implement provisions specified in a waiver application. Waivers are approved for a specified time and can be renewed. CMS may also rescind a waiver at any time for a valid reason. States must demonstrate that the cost of services provided through a waiver does not exceed the costs that would have been incurred without the waiver (a requirement often described as “cost neutrality”). Waiver savings can be used to expand eligibility or offer services that are not otherwise covered under the state’s Medicaid plan. An enrollee can be covered under more than one waiver at the same time.

There are over 30 authorities for different types of waivers [14]. Some of the more frequently used waiver types are described below:

- **Demonstration waivers (Section 1115)** – This authority is for experimental, pilot, or demonstration projects that promote the objectives of the Medicaid and CHIP programs. They give states flexibility to redesign their programs, make various improvements, show the value of innovations, and evaluate policy approaches such as: expand eligibility to individuals not otherwise eligible for Medicaid or CHIP, provide services not typically covered, and use innovative service delivery systems that improve care, increase efficiency, and reduce cost. Demonstration waivers focus on various issues, such as disaster-related services, family planning, substance abuse, premium assistance, enrollee engagement, managed long-term care, services for former foster care youth, and delivery system reform [15].
- **Comparability, Statewideness and Freedom of Choice (Section 1915b)** – This authority allows CMS to waive statutory requirements for comparability, statewideness, and freedom of choice.
- **Home- and Community-Based Services (Section 1915c)** – This authority allows states to provide home- and community-based services (HCBS) as an alternative to institutional services for those individuals who qualify for Medicaid-reimbursable institutional services [16]. There are different types of HCBS waivers including:
  - Individuals over age 65 and individuals with disabilities
  - Physical or intellectual disabilities
  - Intellectual and/or developmental disabilities
  - Brain injury
  - Human Immunodeficiency Virus (HIV)/Acquired Immunodeficiency Syndrome (AIDS)
  - Technology dependent or medically fragile
  - Autism/autism spectrum disorder
  - Mental illness, age 18 and older
  - Mental illness, under age 18
  - Combination with 1115 or 1915(b)
  - Other unspecified populations

State Plan Amendments (SPAs) – SPAs allow states to make certain changes to their state Medicaid plan without requesting approval of a waiver. Various SPAs available to states are identified in the Social Security Act Sections 1915(g) coverage of case management services, 1915(i) home- and community-based services for enrollees under age 65 for mental health and substance abuse disorder services, 1915(j) self-directed personal assistance services, 1915(k) person-centered home- and community-based attendant services and supports - known as the “Community First Choice Option”, 1915(l) coverage for certain enrollees who are patients in certain Institutions for Mental Disease, 1915(a) and 1932(a) voluntary managed care, and 1937(a) benchmark or benchmark equivalent coverage for specific enrollee groups.

2.2.2 Children’s Health Insurance Program (CHIP)

Enacted in 1997 as Title XXI of the Social Security Act, CHIP is also a federal and state partnership. The goal of CHIP is to provide health insurance coverage to low-income children who do not qualify for Medicaid [17]. CHIP provides coverage for children who meet the following criteria [18]:

- Individuals under age 19
- Individuals uninsured and determined ineligible for Medicaid
- Citizens or individuals who meet immigration requirements
- State residents within the state's CHIP income range, based on family income, and any other state-specific rules in the CHIP state plan

States must enroll children in Medicaid, rather than CHIP, if they are eligible for Medicaid. Children may move between Medicaid and CHIP as income and family circumstances change.

Like Medicaid, CHIP is jointly financed with federal, state, and local funds and states must develop and maintain a CHIP (Title XXI) state plan to assure that the state will abide by federal requirements for administering its program and claiming federal matching funds. Unlike Medicaid, the federal government caps the amount of matching funds for CHIP. States have three options for establishing CHIP programs [19]:

- **State CHIP (S-CHIP)** – A program under which a state receives federal funding to provide health assistance to uninsured, low-income children. S-CHIP programs must provide a package of covered services that meet a predefined minimum actuarial standard, but these programs are not required to offer coverage comparable to Medicaid coverage. S-CHIP program specifics vary from state to state.
- **Medicaid expansion CHIP (M-CHIP)** – A program under which a state receives federal funding to expand Medicaid eligibility to targeted low-income children.
- **Combination CHIP** – A program under which a state receives federal funding to implement an M-CHIP program for some children and an S-CHIP program for other children.

As noted above, CHIP is a smaller program than Medicaid, providing coverage for 6.9 million enrollees in November 2021[9]. The program accounted for $18.8 billion in expenditures in federal fiscal year 2019 [20]. The federal share of CHIP spending was 94.2 percent while state and local spending accounted for the remaining 5.8 percent [21].

CHIP covered services include [22]:

- Inpatient and outpatient hospital services
- Physicians, surgical and medical services
- Lab and x-ray
- Well-baby and well-child services, including age-appropriate immunizations
- Mental health and substance abuse disorder services
- Prescription drugs
- Vision and hearing services
- Dental services to promote oral health, restore oral structures to health and function and treatment for emergency conditions
- Other services, optional for states

[9] CHIP enrollees include children covered by S-CHIP, M-CHIP, or Combination CHIP.

2.2.3 T-MSIS Reporting by States

States submit data to CMS monthly on Medicaid and CHIP enrollment, service use, and payments. States extract the data from their operating systems (primarily Medicaid Management Information Systems), recode the data to T-MSIS standards, and submit the data to CMS. CMS and states partner to resolve known data quality issues in their data submissions, although the timeframe for resubmitting data can vary. T-MSIS data are derived from administrative data that are created for program administration purposes, such as enrolling individuals, adjudicating and paying claims, certifying and enrolling providers, assuring fiscal integrity, assessing quality, and performing other management functions. Any data included in T-MSIS are subject to potential data quality issues, although data required for operational purposes are generally more reliable. CMS publishes a Medicaid data quality assessment resource known as the Medicaid and CHIP Data Quality (DQ) Atlas (accessed June 10, 2022) which provides information on the quality of state reported Medicaid data by topic area and state.

2.2.3.1 T-MSIS Analytic Files (TAFs)

There are five Medicaid/CHIP files available to analysts who have access to the linked NHCS-CMS T-MSIS data. The Demographic and Eligibility (DE) file contains demographic and enrollment information on persons enrolled in Medicaid and/or CHIP. The remaining TAFs contain claims records for services provided under fee-for-service (FFS), premium payments to prepaid managed care plans, and encounters for services provided by managed care plans, and are organized into four claims files: inpatient hospital services (IP), long-term care services (LT), pharmacy services (RX), and all other services (OT) organized by date of service. Each of the TAF claims files is organized into separate data files for header, line, and occurrence records (with the exception of RX claims, which only has header and line file types). For more information on how to link claims between these files, please see Section 5.5.

All five files contain 3 years (2015-2017) of T-MSIS data. Researchers should use the FILE_YEAR4 variable to identify the claim year. In the TAF claims files, original state submitted claims, voids, credits, and debits are resolved to create final action claims [23]. However, for IP and LT services, interim claims submitted for payment have not been combined to create completed stay records. CMS publishes a TAF User Guide (accessed June 10, 2022) to assist analysts in understanding how to analyze TAF research files.

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 (DUAL_ELGBL_CD-01 to DUAL_ELGBL_CD_12, by month in the DE), 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). Since Medicaid eligibility is determined for a case (and not a family or household), analysts should use data elements such as household size (T-MSIS data element HSEHLD_SIZE_CD) with caution.

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, but there is no detail on the specific pharmacy services that were provided. 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.

The IP File includes Diagnosis Related Group (DRG) codes which are used to reimburse inpatient hospital services in many states [24]. For DRGs reported in IP claims (T-MSIS data element DRG_CD), the DRG may not be the same as a Medicare DRG. States may use different DRG systems and case weights for Medicaid DRG pricing. Refer to the DRG Code System/Nomenclature variable (T-MSIS data element DRG_CD_SYS and DRG_DESC) for more information on DRGs.

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, which has historically been the case in only a small number of states [25]. 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.

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 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. However, there is a growing trend for immunizations, such as influenza immunizations, to be administered at free-standing pharmacies. 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.[26]

Medicaid payment amounts for prescribed drugs are reported prior to the receipt of manufacturer rebates. Pharmacy records include National Drug Codes (NDC), but for research on prescription drug use, an NDC [27] 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.[10] Analysts should identify any external sources of information to be used in their analyses in their Research Data Center (RDC) proposal (see Section 5.0 for additional information).

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, EPSDT, home health, dental, therapy, transportation, case management, family planning

[10] Drug groupers are available from Wolters Kluwer Health, known as Medi-Span, and First Data Bank.

services[11], 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.

[11] For more details on coverage of family planning services by states see https://www.kff.org/womens-health-policy/report/medicaid-coverage-of-family-planning-benefits-results-from-a-state-survey/ (accessed June 10, 2022)

3 Linkage Methodology

3.1 Linkage Eligibility Determination

The linkage of NHCS patient records to CMS T-MSIS data was conducted through an agreement between NCHS and CMS. Approval for the linkage of patient data with T-MSIS data was provided by NCHS' Research Ethics Review Board (ERB).[12]

Linkage was attempted only for NHCS patient records that had at least two of the following three identifiers present: valid SSN[13], valid date of birth (month, day, and year)[14] or valid name (first, middle initial, and last)[15]. For example, if the PII on the NHCS patient record had no SSN, a full name, and only the year of birth, the record would be considered ineligible for linkage, as only one of the criteria (i.e., that for name) was met.

The variable ELIGSTAT, included on the linked NHCS-T-MSIS match file, provides the linkage eligibility status (which indicates whether the linkage eligibility criteria had been met) for each NHCS patient record. ELIGSTAT values include 0 (ineligible) or 1 (eligible). Table 1 presents the total number of 2016 NHCS patients by age group and sex, the number who were eligible for linkage, the number who were linked at any time in the interval 2015-2017 to CMS T-MSIS claims data, and the percentage of total sample and eligible for linkage who were linked to CMS T-MSIS claims 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 Overview of Linkage

This section outlines steps that were used to link the NHCS data to the CMS T-MSIS enrollment data. For more detailed information on linkage methodology, see Appendix I.

Linkage-eligible NHCS 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.

The NHCS patient records and the CMS T-MSIS enrollment database were linked using both deterministic and probabilistic approaches. For the probabilistic approach, scoring was conducted according to the Fellegi-Sunter method. [28] Following this, a selection process was implemented with the goal of selecting pairs believed to match (i.e., representing the same individual between the data sources).

1. Deterministic linkage joined records on exact SSN, with links validated by comparing other identifying fields (i.e., first name, last name, day of birth, etc.)

[12] 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.

[13] SSN is considered valid if: 9-digits in length containing only numbers, does not begin with 000, 666, or any values after 899, all 9-digits cannot be the same (i.e., 111111111, etc.), middle two and last 4-digits cannot be 0's (i.e., xxx-00-xxxx or xxx-xx-0000), and cannot be 012345678 or 876543210

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

[15] 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, and last) are non-missing.

2. Probabilistic linkage identified likely matches, or links, between all records. All records were probabilistically linked and scored as follows:
    a. Formed pairs via blocking
    b. Scored pairs
    c. Modeled probability – assigned estimated probability that pairs are matches
3. Pairs were selected that were believed to represent the same individual between data sources (i.e., they are a match). Deterministic matches (from step 1) were assigned a match probability of 1 and records selected from the probabilistic match (step 2) were assigned the modeled match probability.

For each NHCS patient-level record that was linked, CMS extracted the T-MSIS claims information and sent the data to NCHS following secure data transfer procedures. Table 1 highlights the linkage results.

Table 1. Linked 2016 NHCS-CMS T-MSIS Claims Records: Sample Sizes and Percent Linked, by Age and Sex

	Sample Size			Percent Linked	
	Total Sample	Eligible for Linkage[3]	Linked to CMS T-MSIS Claims Data[4]	Total Sample[5]	Eligible Sample[6]
2016 NHCS					
Age[1]					
0-17	1,293,458	1,205,473	854,671	66.1	70.9
18-39	1,268,852	1,191,263	671,948	53.0	56.4
40-64	1,130,616	1,062,386	457,400	40.5	43.1
65 and over	762,766	717,624	178,472	23.4	24.9
Total	4,455,692	4,176,746	2,162,491	48.5	51.8
Sex[2]					
Male	2,597,453	1,851,201	915,274	35.2	49.4
Female	3,157,461	2,278,263	1,224,609	38.8	53.8
Total	5,754,914	4,129,464	2,139,883	37.2	51.8

NOTES: Data are presented at patient level.

[1] Age is as of final IP or ED encounter (date of last known contact). Age could not be determined for 1,367,473 patients in the 2016 NHCS due to missing data. Age is calculated by subtracting patient date of birth (DOB) from the final encounter date. When more than one DOB was present, the minimum of the non-missing DOB was selected.

[2] Sex could not be determined for 68,251 patients in the 2016 NHCS due to missing data.

[3] Eligibility for linkage is based upon having sufficient PII in at least two of three data element groups: SSN, name, and date of birth. 1,642,060 patients in the 2016 NHCS were missing all PII and were also considered ineligible for linkage.

[4] This group includes linkage-eligible patients who linked to CMS T-MSIS enrollment database at any time during the linkage interval (2016 NHCS: 2015 – 2017 CMS T-MSIS).

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

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

4 Analytic Considerations

This section summarizes some key analytic considerations for users of the linked NHCS-CMS T-MSIS claims records. It is not an exhaustive list of the analytic concerns that researchers may encounter while using the linked NHCS-T-MSIS data. This document will be updated as additional analytic issues are identified and brought to the attention of the NCHS Data Linkage Team (datalinkage@cdc.gov).

4.1 Analytic Considerations for NHCS Data

4.1.1 NHCS Sampling Weights Are Currently Not Available

Currently, there are no sampling weights available for the 2016 NHCS data. This section will be updated if sampling weights are made available in the future. The hospital-level NHCS sampling approach was conducted with equal probability within strata, which were defined by hospital type, bed size, and urban/rural designation. Unweighted estimates may have bias towards types of hospitals responding at higher rates. These biases will be more of a concern if estimates vary by factors correlated with sampling and response rates.

One way to mitigate these biases in the absence of survey weights is to calculate estimates in the framework of regression modeling that controls for hospital characteristics. This would be done by including hospital characteristics (region, ownership type, and size) as well as patient characteristics (age and sex) among the predictor variables in the model definition. Statistical testing can then be conducted on parameter estimates associated with these characteristics.

4.1.2 NHCS Patient Identification Number

Each patient in the NHCS is assigned a unique identification number, PATIENT_ID. PATIENT_ID does not contain any identifiable information about the patient and is intended to be unique for each individual receiving IP, ED, or OPD services at a participating hospital. However, the de-duplication of patient records required to generate this ID depends on sometimes incomplete or erroneous data, and there may be instances where the same individual is represented by more than one PATIENT_ID. This happens infrequently and should not greatly impact analyses.[16]

4.2 Analytic Considerations for CMS T-MSIS Data Files

4.2.1 Medicaid Data Prior to T-MSIS

CMS began working with states in 2011 to transform the way states report Medicaid data to CMS from the existing national Medicaid Statistical Information System (MSIS) to a new system called Transformed MSIS, or T-MSIS, to improve access to high-quality, timely Medicaid and CHIP data to ensure robust monitoring and oversight of these vital health insurance programs.

The conversion from reporting Medicaid data as MSIS submissions to T-MSIS submissions occurred at different times for each state between 2011 and 2015. CMS produced T-MSIS TAFs

[16] For more information on Patient_ID generation, see Technical Notes on page 14: https://www.cdc.gov/nchs/data/nhsr/nhsr097.pdf (accessed June 10, 2022)

for 19 transitioned states for calendar year 2014 and 30 transitioned states (including DC) in 2015.

Researchers should be aware that the linked 2016 NHCS-CMS T-MSIS data files do not include state-submitted 2015 Medicaid data for the following 21 states: AR, CA, CT, GA, ID, IA, LA, MI, MN, MS, MO, NJ, NY, OR, PA, SD, TN, UT, VT, WV and WY. The linked 2016 NHCS-CMS T-MSIS files include state submitted T-MSIS data for all 50 states plus DC for 2016 and 2017.

4.2.2 State Differences in Medicaid

Though Medicaid is administered under general federal guidelines, there is substantial variation in Medicaid and CHIP programs 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 is substantial variation in the quality of T-MSIS data across states and within a state over time. Consideration of these differences by state may be necessary for analyses that may be affected by these factors. The T-MSIS data element SUBMTG_STATE_CD should be specifically requested in the analyst's RDC proposal if the analyst wishes to incorporate state program characteristics in their analyses. However, although analysts may incorporate state level program characteristics in their analyses, due to disclosure concerns they may not be able to publish state-level estimates. Requests for these types of analyses will be assessed through the NCHS RDC approval process.

4.2.3 Determining Medicaid Program Enrollment

Because Medicaid and CHIP enrollment is linked to income standards, individuals begin and end enrollment in Medicaid and CHIP as income and family situations change. As an individual's eligibility changes, they may enroll and disenroll in Medicaid and/or CHIP throughout the calendar year. This phenomenon is known as "churning" among enrolled populations. "Churning" has important implications for diverse types of research and policy analysis in which analysts use population-based rates. Because of this, analysts may wish to use one or more of the following enrollment definitions to count enrollees and to use as rate denominators for different types of analysis:

- Ever enrolled during the calendar year – The total number of individuals who were enrolled in Medicaid/CHIP at any time of the year, regardless of the length of enrollment.
- Enrolled at a point in time – The number of individuals who were Medicaid/CHIP enrolled on a specific date, often July 1.
- Person-year equivalent enrollment (PYE) – A constructed measure of program enrollment for service use and payment rate analysis, where a person enrolled for three months of the year counts as 25 percent of a PYE enrollee (3/12) and a person enrolled for eight months in the year counts as 67 percent of a PYE enrollee (8/12). This measure adjusts for an enrollee's exposure or risk in the program for use of services and payment for those services.

Data analysts may wish to consider whether certain Medicaid subpopulations should be included in rate denominators. For example, analysts may wish to exclude enrollees with restricted benefits depending on their analytic plan (see Section 4.2.5 for more information on restricted benefit enrollees).

4.2.4 Identifying CHIP Enrollees

CHIP enrollees can be identified and distinguished from other enrollees in the DE file, by month in the calendar year, by using T-MSIS data elements CHIP_CD_1 to CHIP_CD_12. Value = 1 identifies individuals who were not enrolled in either an M-CHIP or an S-CHIP program. Value = 2 identifies individuals who were enrolled in an M-CHIP program. Value = 3 identifies individuals who were enrolled in S-CHIP. Value = 4 identifies individuals who were enrolled in both M-CHIP and S-CHIP. Analysts should interpret value = 4 as indicating that an individual was enrolled in S-CHIP for part of the month and M-CHIP for a different part of the same month.

4.2.5 Identifying Restricted Benefit Enrollees

States have the option to limit certain enrollees to a set of restricted benefits including limiting Medicaid covered services to only family planning, pregnancy care, or substance use disorder treatment. Although all states are currently required to include some form of covered family planning services for their Medicaid enrollees, states also have the option to enroll certain individuals whose Medicaid coverage is limited to the use of family planning services only. Many states have enrolled individuals under this restricted benefits option, but for most states the number of these enrollees is small [29]. However, California has extended eligibility to over 1 million individuals under this option [30]. Some states do not collect and report enrollee demographics for restricted benefit individuals. Information on specific restricted benefits enrollment is available by month on the DE file in variables RSTRCTD_BNFTS_CD_01 through RSTRCTD_BNFTS_CD_12. Researchers should consider whether it is appropriate to remove restricted benefit enrollees from their specific analyses as they are only eligible for those specific Medicaid covered services.

4.2.6 Dually Eligible Individuals

Certain individuals, known as dually eligible individuals, are enrolled in both Medicare and Medicaid. In 2018, there were 12.2 million dually eligible individuals, including persons over age 65 and persons with disabilities. For these individuals, Medicare is the first payer for services covered by Medicare Parts A, B, C, and D. Medicaid provides supplemental coverage for covered Medicare services including copayment and deductible amounts up to the limits identified in the state Medicaid plan and also pays for Medicare Part B premiums for all dually eligible individuals. Most dually eligible individuals receive full Medicaid benefits, but some dually eligible individuals receive only restricted benefits (known as partial benefits). Partial benefit dually eligible individuals typically represent less than 10 percent of all dually eligible individuals, but percentages vary by state [31]. For those dually eligible individuals who receive full Medicaid benefits, Medicaid typically covers institutional long-term care services, Medicaid covered drugs and other pharmacy-dispensed items beyond the scope of Medicare Part D coverage, and other services such as transportation and various types of therapy not generally covered by Medicare [32]. For crossover claims, for which both Medicare and Medicaid pay the same provider for services covered under each program, the Medicaid claim may be missing some important details, such as diagnoses and procedures, which can be found on the Medicare claims. Medicaid data provide a limited view of total use and cost for dually eligible individuals. Analysts using linked T-MSIS data files should consider if and how they want to include dually eligible individuals in their analyses. To obtain a more complete view of services and costs for dually eligible individuals, analysts may wish to consider analyzing both linked Medicaid and Medicare claims data for linked 2016 NHCS patients (see Section 4.3.2 for additional information).

Different categories of dually eligible individuals can be identified in the Demographic and Eligibility (DE) file, by the values presented in Table 2.

Table 2: T-MSIS Code Values for Dually Eligible Individuals

Dually Eligible Individuals Groups	Monthly T-MSIS DUAL_ELGBL_CD Values
Full Benefit Dually Eligible Individuals	
Qualified Medicare Beneficiaries (QMB Plus)	02
Specified Low-income Medicare Beneficiaries (SLMB Plus)	04
Restricted Benefit Dually Eligible Individuals	
QMBs – only	01
SLMBs – only	03
Qualified Working Disabled Individuals (QDWIs)	05
Qualifying Individuals (QIs)	06
S-CHIP Enrollees Entitled to Medicare	10

4.2.7 Managed Care

Health care through Medicaid (and CHIP) is delivered through FFS and managed care programs. Medicaid managed care programs are insurance plans in which a health care organization provides a defined bundle of health services for a fixed monthly fee paid by the state's Medicaid program. States use an array of different types of managed care arrangements in Medicaid. Medicaid managed care plans include comprehensive plans that cover most (but not necessarily all) enrollee health services. Other plans provide coverage for limited services, and service coverage can vary by plan type. Since the 1990s, state Medicaid programs have increasingly relied on managed care to organize and deliver services. The percentage of Medicaid enrollees who are served by managed care plans has increased steadily in recent years [33]. There are 21 types of managed care plan types that states can choose as part of their state plan. These types are organized into the following higher-level categories:

- **Comprehensive Managed Care Organizations (MCOs) –** These plans provide acute, primary and specialty services. Some plans include behavioral health and long-term care services and supports. Examples: Health Maintenance Organizations (HMOs) and Health Insuring Organizations (HIOs).
- **Prepaid Ambulatory Health Plans (PAHPs)** – These plans provide ambulatory services (e.g., transportation) but they do not arrange for or have responsibility for the provision of inpatient hospital or institutional services.
- **Prepaid Inpatient Health Plans (PIHPs)** – These plans are responsible for the provision of inpatient hospital or institutional services and often cover behavioral health and intellectual/developmental disabilities and support services.
- **Program for All-inclusive Care for the Elderly (PACE)** – These plans provide comprehensive medical and social services in an adult day care center as well as in-home and referral services, as needed.
- **Other types of Prepaid Health Plans (PHPs)** – For example, such plans may cover dental care or long-term care services and supports.
- **Primary Care Case Management (PCCM)** – These plans assign an enrollee to a primary care provider who oversees and coordinates the enrollee's care.

For all types of managed care plans, except PCCMs, a state pays plans a monthly premium, the plans provide care to enrollees, and there is no additional payment by Medicaid. PCCM providers manage an enrollee's care but do not receive a prepaid premium. States typically pay PCCM providers a monthly FFS payment to manage an enrollee's care. States must obtain a waiver of the freedom of choice requirement from CMS to require enrollees to join managed care plans.

As of July 1, 2019, states covered 65.7 million enrollees (83.5 percent of total enrollment) in some form of managed care [34]. However, analysts should not assume that all individuals enrolled in a managed care plan receive all Medicaid covered services as part of their managed care plan. Enrollees can be covered in a non-comprehensive managed care plan for some services and FFS for other Medicaid services. Even comprehensive plans may have 'carve outs' for some services, such as prescribed drugs and dental services which are not covered by the managed care plan. Furthermore, an enrollee can be enrolled in more than one type of managed care plan at the same time. For example, an enrollee could be simultaneously enrolled in three managed care plans: dental, behavioral health, and pregnancy related services.

There is variation in the extent of managed care enrollment overall and by type of plan across the states and the District of Columbia. For example, as of July 1, 2019, Alaska and Wyoming had little to no managed care enrollment. Conversely, Hawaii, Nebraska, and Puerto Rico covered over 95 percent of their enrollees in comprehensive managed care plans. Seven states provided coverage for over 50 percent of their enrollees in PCCMs [35].

Since it is possible for an individual to be enrolled in more than one type of managed care plan (as well as FFS) at any point in time, the DE file identifies up to 12 managed care types, per month, in T-MSIS data elements MC_PLAN_TYPE_CD_01 to MC_PLAN_TYPE_CD_12. The DE base record does not include managed care plan identifiers, so it is not possible to identify the specific plans in which the individual was enrolled.

For services provided through managed care plans, encounter reporting lags behind FFS claims reporting, but for many plans, it is fairly complete by the time that TAFs are produced. However, CMS's ability to establish adequate benchmarks for encounter reporting is limited, so data quality issues may exist. Medicaid payment amount (T-MSIS data element MDCD_PD_AMT) should be $0 for encounter records, but if payment amount is greater than $0, those amounts should be disregarded for analytic purposes as a state pays a monthly premium payment to the plan instead of reimbursing individual claims. Premium payments to plans are identified in T-MSIS data element CLM_TYPE_CD, value = 2 for Medicaid and value = B for CHIP. Encounter records for services provided under prepaid managed care plans are also identified in data element CLM_TYPE_CD, value = 3 for Medicaid and value = C for CHIP.

4.2.8 Waiver and Demonstration Reporting

Individuals can be enrolled in various state waivers. The DE base record does not include any information on waiver enrollment, but the TAF claims files (IP, LT, OT, RX) include data elements to identify services provided under waivers. Values of T-MSIS data element WVR_TYPE_CD identify the type of waiver under which a service was provided and T-MSIS data element WVR_ID is the state-assigned identifier for the waiver. Researchers interested in learning more about 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, to obtain more detailed information on individual waivers.

4.2.9 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 significantly by state. An example of a service tracking claim is a claim for a nursing home per diem rate adjustment that applies to all Medicaid covered residents of the facility at a particular time. Because service tracking claims cannot be linked to an individual, they have been excluded from the linked 2016 NHCS-T-MSIS files.

4.2.10 Missing Enrollment Data and Dummy Enrollment Records

There are instances in which there are valid claim records for an enrollee, but there is no associated state-reported enrollment record. CMS has created ‘dummy’ enrollment records for these enrollees in the Demographic and Eligibility (DE) file. Analysts will need to determine if they want to include dummy enrollment and their associated claims records in their analyses as there is typically no available demographic information for these enrollment records. DE ‘dummy’ records can be identified using the T-MSIS DE data element MISG_ELGBLTY_DATA_IND, code value = 1. Some dummy enrollment records may include demographic data if they were linked to enrollment records in other years.

4.2.11 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 (accessed June 10, 2022) includes an analysis of payment consistency between header and line-item claims records for the four claims file types [36]. The TAF claims files (IP, LT, RX, and OT) include a variable (PYMT_LVL_IND) that indicates whether the Medicaid claim payment was made at the header or line-item level. Analysts should use caution when analyzing type of service (TOS) codes in line-item payment claims[17]. The quality of TOS reporting is still under CMS review due to the substantial increase in the number of TOS categories available in T-MSIS (compared to prior Medicaid data reporting requirements), the lack of clear definitions in the Code of Federal Regulations (CFR) for certain T-MSIS TOS categories, and the potential for inconsistencies in state mapping of existing TOS categories to the new T-MSIS TOS values. This data element cannot be compared meaningfully across states.

4.2.12 Mother and Newborn Claims Records

States use different methods to report labor and delivery services provided to women and their newborns. Some providers 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. Delivery services provided to the mother, and services provided to the newborn may also be included in a single claim[18]. For example, a mother may have a two-day hospital stay and the newborn may have a one-day stay, both being discharged on the second day. In

[17] TOS is available only in line-item claims.

[18] See frequently asked questions #2437, #2463, and #3557 at https://www.cms.gov/files/document/frequently-asked-questions-9 (accessed June 10, 2022).

this example, the length of stay may be reported as three days (two for the mother and one for the newborn). It is also possible that a mother and her newborn may share the same MSIS identifier [37].

4.2.13 Claims Reporting Lags

In certain circumstances there may be delays in claims reporting for pregnant women who apply for Medicaid after they become pregnant because states may choose to provide Medicaid coverage retrospectively to the beginning of the pregnancy.

States may also delay reporting claims for certain health services until the claim payment adjudication process is completed. There may be variation in claims reporting lags by state and by claim type. The 2015-2017 T-MSIS TAFs linked to the 2016 NHCS were finalized by CMS in 2020 [38].

4.2.14 Multiple Claims with the Same Service Date

Due to the manner in which health care claims are submitted for reimbursement for certain TOS, 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 service provided billed separately.

4.2.15 General Limitations of Medicaid Data

There are general limitations to the information contained in the T-MSIS files. Because these files contain only Medicaid-paid services, they do not capture service use or payments during periods of non-enrollment, services paid by other payers, or services provided at no charge. Because T-MSIS files consist only of enrollee-level information, they do not include prescription drug rebates received by Medicaid, aggregate Medicaid payments made to disproportionate share hospitals (DSH) (hospitals that serve a disproportionate share of low-income patients with special needs), payments made through upper payment limit (UPL) programs, and payments to states to cover administrative costs.

4.2.16 T-MSIS Data Quality

With any new data system, there are data quality issues and concerns in the early years after implementation. CMS instituted a continuing process of quality improvement for T-MSIS by identifying a series of T-MSIS Priority Items (TPIs), as follows:

- An initial list of 12 highest TPIs identified in 2017 [39]
- The list expanded to a total of 23 items in 2019 [40]
- The list was again expanded to a total of 32 items in 2020 [41]

State progress in addressing TPIs 1-23, as of July 2021 is available at references listed above.

4.2.16.1 Guidance Documents on T-MSIS Data Quality

CMS has produced a data quality assessment resource known as the Medicaid and CHIP DQ Atlas (accessed June 10, 2022) for each public release of annual TAF data. This resource enables data users to examine the data quality for enrollment, claims, service use, and payment data. The DQ Atlas (accessed June 10, 2022) is searchable by data topic and by state. For each state, DQ assessments assign one of five values to indicate the extent to which T-MSIS data elements are usable, reliable, and accurate for analyzing the selected topic, based on comparisons to expected data patterns or external data benchmarks.

Subject matter topical areas discussed in the DQ Atlas (accessed June 10, 2022) include:

- Enrollment benchmarking
- Enrollment patterns over time
- Enrollee information
- Claims files completeness
- Expenditure benchmarking
- Payments
- Service use information
- Provider information
- Non-claims records

Specific data quality issues are also available for each of the four TAF claims file types (IP, LT, RX, and OT).

The State Health Access Data Assistance Center (SHADAC) has produced an analysis of the quality of race and ethnicity data reported in the 2018 T-MSIS data [42]. The Research Data Assistance Center (ResDAC) is also a valuable resource for information on T-MSIS data [43].

4.2.16.2 Guidance on Analyzing Illinois Claims Data

There are special reporting issues that apply to T-MSIS claims data from the state of Illinois. Since analysts will not be aware if the patient is from the state of Illinois, they should consider requesting the variable SUBMTG_STATE_CD in their RDC proposal to determine if their analysis will include Illinois claims data. For details, on how to handle these records see TAF Technical Guidance: How to Use Illinois Claims Data (accessed June 10, 2022).

4.3 Analytic Considerations for Linked NHCS - CMS T-MSIS Data Files

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

NHCS patients may have multiple DE records. Most often, this is because a patient is linked to several years of T-MSIS data. However, a patient may be linked to multiple DE records within the same year. There are multiple explanations for this situation including Medicaid enrollees enrolling in Medicaid in more than one state as they move between states (i.e., there will be DE records for each state in which an individual is enrolled), eligibility changes resulting in survey participants dis-enrolling and re-enrolling in Medicaid within the same year if the state did not retain the same Medicaid identification number for that enrollee, and errors in administrative data systems or linkage methodology. Most NHCS patients with multiple DE records per year had Medicaid enrollment in more than one state.

The existence of multiple DE records within a given year with overlapping months of Medicaid enrollment data between the DE records can complicate analyses. It is possible for an individual to be enrolled in one state for part of the month and another state during the same month. Also, it is possible for an individual to be enrolled in more than one state at the same time as there is no requirement for individuals to terminate enrollment if they move to a different state. In considering how to assess Medicaid enrollment in the presence of multiple DE records within a year, analysts may consider the use of data elements that indicate enrollment by month in each record. The data elements ELGBLTY_GRP_CD_1 to ELGBLTY_GRP_CD_12 can be analyzed across multiple DE records to create a summary of Medicaid enrollment across all months within a given year.

4.3.2 Payments for Medicare Covered Services for Dually Eligible Individuals

Because Medicare is the primary payer for Medicare covered services for dually eligible individuals, much of these individuals' health care cost and utilization data will be found in the linked 2016 NHCS-CMS Medicare files. Service utilization records for services covered by Medicaid and not Medicare will be found in the linked 2016 NHCS-CMS T-MSIS data files.

Beginning in 2006, dually eligible individuals began receiving the Medicare Part D drug benefit, and their utilization for Part D-covered drugs is provided in the linked Medicare Part D Event data. Medicaid has the option to cover drugs not covered by Medicare. Records for those drugs will be included in the linked 2016 NHCS-CMS T-MSIS data files. The 2016 NHCS is also linked to the 2016–2017 Medicare Part D Prescription Drug Event data files.

Analysts interested in analyzing linked Medicare claims and prescription drug data for dually eligible individuals should request to use these linked 2016 NHCS-CMS Medicare files in their RDC proposal. For more information about the linked 2016 NHCS-CMS Medicare data files, see Section 5.2.1.

4.3.3 T-MSIS Match File

The linked T-MSIS Match file can be used to identify which of the NHCS patients were eligible for linkage and linked to a T-MSIS demographic and eligibility record. This file contains one record for each NHCS 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). NHCS 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 NHCS-CMS T-MSIS linkage, NCHS used a probabilistic cut-off value of 0.92 to determine final match status. Candidate pairs with a PROBVALID that exceeded the probabilistic cut-off (i.e., PROBVALID>0.92) were deemed a link. The estimated type I error was 0.04 and the type II error was 1.5%. For additional discussion on 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. In order to do this, researchers should increase the probability cut-off value (to a value closer to 1.0). Of note, the PROBVALID cannot be decreased from 0.92. To change the NCHS link acceptance cut-off value, researchers should request the variable PROBVALID in their RDC proposal (see Section 5).

4.3.4 Temporal Alignment of Survey and Administrative Data

The 2016 NHCS has been linked to 2015-2017 Medicaid T-MSIS data files. Linked Medicaid data may be analyzed in conjunction with patient records collected during the 2016 survey year. However, users should be aware that 2016 NHCS patients may have linked Medicaid data for 2015 only, 2016 only, 2017 only, or they may have linked Medicaid data that spans different intervals from 2015 through 2017 depending on their Medicaid enrollment period. The linked NHCS-CMS T-MSIS data can provide a more complete profile of health care services utilization, including information about patient's use of hospital services at hospitals not participating in the NHCS,   for those NHCS patient records linked to Medicaid data.

4.3.5 Merging NHCS Analytic Files to the Linked NHCS-CMS T-MSIS Data Files

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

The NHCS analytic files include analytically pertinent hospital-level details (such as bed size and geographic region) and episode-level details (patient demographics, diagnoses, procedures, admission and discharge dates). To perform NHCS patient encounter-level analysis, the linked NHCS-CMS T-MSIS data files must be used in conjunction with the 2016 NHCS analytic files.[19] The shared variable, PATIENT_ID, allows analysts to merge NHCS patient records with   the linked NHCS-CMS T-MSIS data files.

4.3.6 Merging Within the Linked NHCS-CMS T-MSIS Data Files

Researchers should use PATIENT_ID, MSIS_SEQN[20], and FILE_YEAR4 to merge enrollment information from the demographic and eligibility base file to each of the TAF claims files.

The linked TAF claims files include separate files for claims header, line item, and occurrence[21]. To merge claim header, line, and occurrence information within a unique claim record, researchers should use PATIENT_ID, MSIS_SEQN, FILE_YEAR4, and NCHS_CLM_ID. For example, researchers who wish to merge claim header and line item information for a specific OT claim record would merge data using the unique combination of PATIENT_ID, MSIS_SEQN, FILE_YEAR4, and NCHS_CLM_ID[22].

5 Access to Data Files

[19] Find more information about the NHCS analytic files: https://www.cdc.gov/rdc/b1datatype/dt1224h.htm (accessed June 10, 2022)

[20] MSIS_SEQN was created by NCHS to mask MSIS identifiers. The MSIS_SEQN represents the combination of MSIS_ID and State_CD

[21] Pharmacy (RX) claims include header and line item information only.

[22] NCHS_CLM_ID was created by NCHS to mask the original T-MSIS claims identification numbers.

5.0 Access to the Restricted-Use Linked NHCS-CMS T-MSIS Claims Data Files

To ensure confidentiality, NCHS provides safeguards including the removal of all personal identifiers from analytic linked files. Additionally, the linked data files are only made available in secure facilities for approved research projects. Researchers who wish to access the linked 2016 NHCS-CMS T-MSIS claims data files must submit a research proposal to the NCHS Research Data Center (RDC) to obtain permission to access the restricted use files. All researchers must submit a research proposal to determine if their projects are feasible and to gain access to these restricted data files. The proposal provides a framework which allows RDC staff to identify potential disclosure risks. More information regarding the RDC and instructions for submitting an RDC proposal are available from: https://www.cdc.gov/rdc/ (accessed June 10, 2022).

5.1 Variables to Request in RDC Proposals

The restricted-use variables from the 2016 NHCS (accessed June 10, 2022) and the variables of interest from the linked 2016 NHCS-CMS T-MSIS files must be specifically requested as part of a researcher's proposal to the RDC. Staff in the RDC verify the full list of variables and check for potential disclosure risk.

A complete set of codebooks, providing information on the variables for each of the T-MSIS TAFs, has been created to assist researchers in the variable selection process. There is a single codebook for each TAF (DE, IP, OT, LT, and RX) that combines the variables from each of the claim file types (header, line, and occurrence). Using the inpatient claims files as an example, the variables in the header will appear first in the codebook, immediately followed by the variables in the inpatient line file, and finally the variables in the inpatient occurrence file. A column has been added to the codebook indicating which data file the variable is associated with. Note that researchers must specify the TAF and the claim file type (header, line, or occurrence) for each requested variable in the RDC proposal.

To obtain information on 2016 NHCS patient eligibility for linkage and CMS T-MSIS match status, researchers should request access to the 2016 NHCS-CMS T-MSIS Match Status file. (See Section 4.3.3 for more information regarding the variables available on the CMS T-MSIS Match Status file).

To incorporate state-level differences in Medicaid program characteristics, researchers will need to request the variable SUBMTG_STATE_CD. Analysts can incorporate state-level data into their analyses of 2016 NHCS-CMS T-MSIS linked data but may not be allowed to remove analytic results that include specific state codes from the RDC. Researchers interested in incorporating state-level Medicaid program characteristics in their analysis should provide information in their RDC proposal about how they intend to publish their results.

Analysts proposing to analyze 2016 NHCS-CMS T-MSIS linked claims data should request access to the Demographic and Eligibility (DE) TAF for the same calendar years as the Medicaid health care claims TAFs (IP, LT, RX, OT) in order to determine the correct study denominators for the linked Medicaid population. (See Section 4.2.3 for more information regarding determining Medicaid program enrollment).

Researchers must request both NHCS and T-MSIS identification numbers in order to merge variables between the NHCS analytic files and the 2016 NHCS-CMS T-MSIS linked claims data and when merging claims within the 2016 NHCS-CMS T-MSIS linked claims. Please see Sections 4.3.5 and 4.3.6 for more information on which identification numbers to request in your RDC proposal.

5.2 Additional Related Data Sources

5.2.1 Linked NHCS-CMS Medicare Files

Analysts interested in studying health care utilization and costs for the dually eligible population (persons enrolled in both Medicare and Medicaid) may wish to also request access to the 2016 NHCS-CMS Medicare linked data files (accessed June 10, 2022) for enrollment and claims data from 2016–2017. Medicare is the first payer for health care services covered by Medicare Parts A, B, C, and D, with Medicaid providing supplemental coverage for covered Medicare services including copayment and deductible amounts up to the limits identified in the state Medicaid plan. (See Section 4.2.6 for more information regarding health care claims processes for dually eligible individuals).

The linked 2016 NHCS-CMS Medicare Master Beneficiary Summary Files (MBSF) include information on Medicare program entitlement and enrollment, summarized annual health care utilization and cost data, and chronic condition flags indicating the presence of certain health conditions for linked Medicare beneficiaries. Additionally, the 2016 NHCS-CMS Medicare linked data files include health care claims and encounters, prescription drug events, and patient assessment data for linked Medicare beneficiaries. To integrate the linked NHCS-CMS Medicare linked data files into the linked NHCS-CMS T-MSIS Claims Data files, joins are made on the common identification number, PATIENT_ID.

More information about the linked 2016 NHCS-CMS Medicare data files can be found at: https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm (accessed June 10, 2022).

5.2.2 Linked NHCS-NDI Mortality Files

Analysts interested in studying mortality among the 2016 NHCS patient population enrolled in Medicaid are encouraged to use the linked mortality data available in the 2016 NHCS- NDI Mortality files rather than the mortality data available in the linked 2016 NHCS-CMS T-MSIS files. The linked 2016 NHCS-NDI Mortality files (accessed June 10, 2022) include information on deaths identified for the entire 2016 NHCS patient population through linkage with the National Death Index and are not limited to deaths among the Medicaid enrolled population. In addition, in the NHCS-NDI linked data cause of death is available for patients who died. The linked mortality file includes Patient ID, date of birth, date of death, and cause of death information for linked decedents. To integrate the linked NHCS-NDI linked data files into the linked NHCS-CMS T-MSIS data files, joins are made on the common identification number, PATIENT_ID.

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

Researchers interested in outcomes related to housing insecurity may also request variables from the linked 2016 NHCS–2015-2017 HUD Administrative Data file if housing assistance is a variable/outcome of interest (Restricted-Use Linked NHCS – HUD Administrative Housing Data, accessed June 10, 2022). The linked HUD administrative data files include variables pertaining to the recipient's participation in Housing Choice Voucher (HCV), Public Housing (PH), and/or Multifamily (MF) programs. To integrate the linked NHCS–HUD administrative data files into the linked NHCS-CMS T-MSIS data files, joins are made on the common identification number, PATIENT_ID.

Appendix I: Detailed Description of Linkage Methodology

1 NHCS and CMS T-MSIS Linkage Submission Files

Prior to the linkage of the NHCS and CMS T-MSIS administrative records, there were a series of processes that performed various data cleaning routines on the PII fields within each of the files. Of note, processing was conducted separately for NHCS and CMS T-MSIS records. The following PII fields were individually processed and output to its own file (i.e., there were separate files for SSN, DOB, name, etc., each record showing a possible value for that field for each patient (NHCS) or enrollee (CMS T-MSIS)):

- SSN (validated)[23]
- DOB (month, day, and year)
- Sex
- 5-Digit ZIP code and state of residence
- First, middle, and last name

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

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

Similar to the cleaning process, a more elaborate routine was used to generate alternate records involving the name fields. Additional records were generated for patients with multiple name parts, common nicknames, and for common Hispanic and Asian names. NCHS created a common nickname lookup file which was used to generate a second record replacing the nickname with the formal name. Table 3 below provides two examples of how multiple part name information was used to generate alternate records, using hypothetical data. For patient A, the first name was used to generate multiple records, and for patient B, the last name was used.

[23] SSN is considered valid if: 9-digits in length, containing only numbers, does not begin with 000, 666, or any values after 899, all 9-digits cannot be the same (i.e., 111111111, etc.), middle two and last 4-digits cannot be 0’s (i.e., xxx-00-xxxx or xxx-xx-0000), and is not 012345678 or 876543210

Table 3. Example of Alternate Record Generation using Name Fields

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

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 for NHCS patient records and for CMS T-MSIS enrollment records, separately. During this process, multiple submission records were created for each patient/enrollee to show all combinations of the recorded values for these fields. That is, if a patient/enrollee 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/enrollee (see Table 4 for example). Submission records that did not meet the eligibility requirements (see Section 3.1 Linkage Eligibility Determination) were removed from the submission file.

Table 4. 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

2 Deterministic Linkage Using Unique Identifiers

The deterministic linkage, which was the next step in the linkage process, used only the NHCS and CMS T-MSIS submission records that included a valid format SSN. The algorithm performed two passes on the data, first checking for full 9-digit SSN agreement and then for records where the last 4-digits of the SSN agreed. After records had been matched 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 1/2 (1^{st} pass using SSN-9) or greater than 2/3 (2^{nd} 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 2^{nd} pass were required to have at least 5 non-missing PII variables in agreement to be deemed a deterministic match. Of note, NHCS 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 collection of records resulting from the deterministic match is referred to as the 'truth source.'

3 Probabilistic Linkage

The second step in the linkage process was to perform the probabilistic linkage. 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. [44] 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 data linkage expert Peter Christen, blocking or indexing, "splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key)." [45] Intuitively developed rules can be used to define the blocking criteria, however, for this linkage, 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' as the validation dataset and a sample of the NHCS and CMS 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." [46], [47]

The machine learning algorithm learned 14 blocking passes to be used in the blocking scheme. Table 5 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 in that blocking pass. Further, if the ZIP code of residence was used as a blocking variable and state of residence was not, then state of residence was excluded from the list of scoring variables as it is implied to be in agreement on all records. Additionally, since sex was found to have minimal contribution as a scoring variable and is highly correlated with first name agreement, sex was not included in the pool of potential scoring variables but was used as a blocking variable.

Table 5. Blocking and scoring scheme used to identify and score potential links

Key Number	Blocking Key	Scoring Key
1	Last name, month of birth, day of birth, year of birth	First name, middle initial, state of residence, ZIP code of residence
2	Month of birth, day of birth, year of birth, state of residence, sex	First name, middle initial, last name, ZIP code of residence
3	Last name, first name, state of residence, sex	Middle initial, month of birth, day of birth, year of birth, ZIP code of residence
4	Last name, month of birth, year of birth, state of residence, sex	First name, middle initial, day of birth, ZIP code of residence
5	First name, month of birth, year of birth, state of residence, sex	Middle initial, last name, day of birth, ZIP code of residence
6	Last name, month of birth, day of birth, state of residence, sex	First name, middle initial, year of birth, ZIP code of residence
7	First name, month of birth, day of birth, state of residence, sex	Middle initial, last name, year of birth, ZIP code of residence
8	Last name, first name, month of birth, year of birth	Middle initial, day of birth, state of residence, ZIP code of residence
9	Day of birth, year of birth, state of residence, ZIP code of residence	First name, middle initial, last name, month of birth
10	Last name, first name, day of birth	Middle initial, month of birth, year of birth, state of residence, ZIP code of residence
11	First name, month of birth, day of birth, year of birth	Middle initial, last name, state of residence, ZIP code of residence
12	Last name, year of birth, state of residence, ZIP code of residence, sex	First name, middle initial, month of birth, day of birth
13	Last name, day of birth, year of birth, state of residence, sex	First name, middle initial, month of birth, ZIP code of residence
14	Month of birth, year of birth, state of residence, ZIP code of residence	First name, middle initial, last name, day of birth

3.2 Score Pairs

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

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

3. Calculate pair weight scores

The pair scores were calculated on the 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)

3.2.1 Calculate M- and U- Probabilities

The **M-probability** – the probability that the identifiers using the records in question agree, given that records represent the same person – were estimated separately within each individual blocking pass. M-probabilities were calculated for each of the identifiers not used in the blocking key (Table 5). Within the blocking pass, pairs with agreeing SSN (defined as 8 or more digits being the same) were used to calculate the M-probabilities, as these are assumed to represent the same individual. 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 record for each patient ID and CMS T-MSIS ID (see Tables 6 and 7 for example of record summarization). For example, among qualifying pairs in blocking pass 2, 99.4% agree on day of birth and 94.5% agreed on state of residence. These percentages represented estimates of the M-probabilities for these identifiers.

Table 6. Example of Agreement Flags for Blocked Records

Person Identifiers		**PII Agreement flags[1]**					
Patient ID	**CMS T-MSIS ID**	**Day of birth**	**Month of birth**	**Year of birth**	**ZIP Code**	**State of residence**	**Sex**
1	1	1	0	1	0	.	1
1	1	.	1	1	0	0	1
1	1	1	0	1	0	0	0
2	2	1	0	1	0	0	0
3	789	1	1	.	0	1	0
3	789	0	1	0	1	1	0
3	789	.	1	0	1	.	1
3	789	0	0	1	1	1	1
3	322	1	0	1	1	1	1

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

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

Table 7. Example Showing Summarization of Blocked Records for M-Probability Estimation

Person Identifiers		PII Agreement flags[1]					
Patient ID	CMS T-MSIS ID	Day of birth	Month of birth	Year of birth	ZIP Code	State of residence	Sex
1	1	1	1	1	0	0	1
2	2	1	0	1	0	0	0
3	789	1	1	1	1	1	1
3	322	1	0	1	1	1	1

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

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

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

- First/last initial agreement – used in the scoring process when only an initial was present in the name field
- 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** – 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 only calculated 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 SSN were not in agreement (defined as having less than 5 matching digits). 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 that had majority of the non-missing PII among scoring variables were in agreement) were excluded prior to calculating the U-probabilities. For example, when computing the U-probability for day of birth in blocking pass 12, records that did not agree on SSN that had majority of the PII among first name, middle initial, and month of birth were excluded from the assumed non-matches. These records were assumed to be probable matches given that a majority of the PII between the survey and administrative records were in agreement.

The U-probabilities, however, were calculated for each value (level) of a variable. For example, the state of residence U-probabilities within blocking pass 1 for Florida and Pennsylvania were, 0.052 (5.2%) and 0.091 (9.1%), respectively. However, for first and last name, the U-probabilities were calculated in a different manner further described in Section 3.2.2.

3.2.2 M- and U- Probabilities for First and Last Names

For first and last name M and U-probabilities, corresponding Jaro-Winkler levels (0.85, 0.90, 0.95, and 1.00) are calculated. The Jaro-Winkler algorithm assigns a string similarity score, between 0 and 1 (both inclusive), depending on the likeness between two strings. For example, if the first name on the survey record were Albert and on the CMS T-MSIS record it was Abert, this would receive a Jaro-Winkler score of 0.96. For M-probabilities, the manner of their creation is identical to the process described above. For example, the M-probability for first name at the Jaro-Winkler 0.90 level is the rate of agreement for all first names with a Jaro-Winkler score of 0.90 and above.

Because of the large number of unique name values, it was impractical to compute U-probabilities specific name for each blocking pass (i.e., there would not be enough records available for it to be done accurately). Instead, U-probabilities were estimated using pairs generated by the Cartesian product of all records in the NHCS submission file and a simple random sample of 3% (6,294,662 records for first name and 6,356,739 records for last name) of records with non-missing name information of the CMS T-MSIS submission file.

Complete name tallies (separately, for first and last names) were then produced for the NHCS submission file. For each level of name on the file, 100,000 names were randomly selected from the CMS T-MSIS submission file 3% sample to compare to it. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. The number of names in agreeance of the 100,000 randomly selected CMS T-MSIS file names that agreed at that level for each name were then tallied. [48], [49], [50]

3.2.3 Calculate Agreement and Non-Agreement Weights

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

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

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

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

3.2.4 Calculate Pair Weight Scores

In the next step, pair weights were calculated 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 follow the same general process:

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

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

3.3 Probability Modeling

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

- Combine pairs across blocking passes (Pair-weights are specific to each blocking pass and are not comparable)
- Select a "best" record among patient's IDs that have linked to multiple administrative records
- Select final matches based on a probability threshold (discussed in the following section)

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

1. A pair-weight adjustment 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, $N_{\widehat{matches,B}}$, 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.

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

   Note that in the first iteration, it was assumed that $N_{\widehat{matches,B}}$ = $N_{\widehat{non-matches,B}}$, resulting in $Adj_B = 0$. If, however, in a later iteration, the number of matches was estimated to be, $N_{\widehat{matches,B}}$ = 20,000, 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*, were computed in blocking pass, *B*, being a match by taking 2 to the power of the adjusted pair-weight (sum of pair-weight (*PW*) and Adj_B, the blocking pass pair weight adjustment).

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

   Continuing with the example from Step 1…

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

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

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

   Continuing with the example…

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

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

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

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

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

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

3.4 Adjustment for SSN Agreement

Up to this point, every pair generated through the probabilistic routine was assigned a value that estimates its probability of being a match. However, this estimate did not take SSN agreement into account. This was conducted as a separate step because for the other comparison variables, M- and U- probabilities were estimated based on probable matches or non matches that were determined based on SSN agreement and clearly this was infeasible for SSN itself.[24]

To remedy this, before the algorithm adjudicated the matches against the probability threshold, one final adjustment was made to the match probabilities (for probabilistic pairs). For pairs that had an SSN on both the NHCS and CMS T-MSIS record, the estimated probability was adjusted based on the last four digits of the SSN.[25]

When the last four digits of SSN[26] 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 NHCS or CMS T-MSIS record. So, for these pairs:

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

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

4.1 Estimating Linkage Error to Determine Probability Cutoff

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

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

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

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

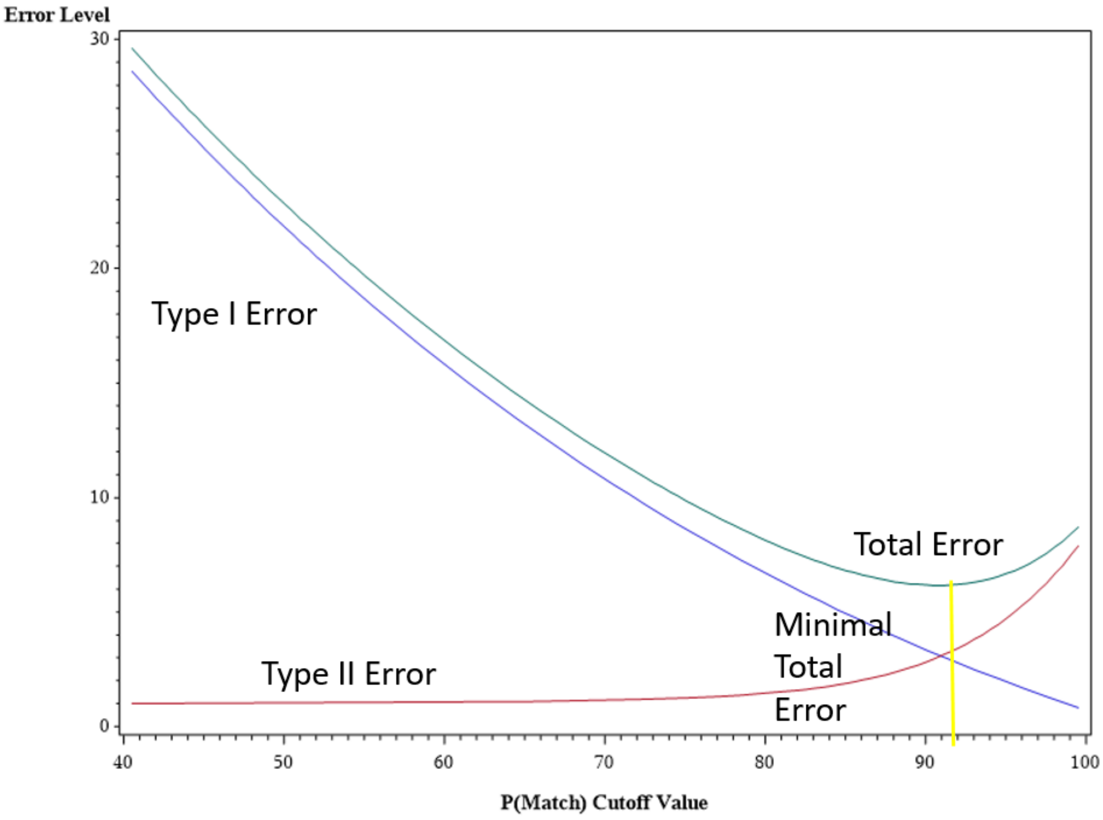
Because all records were included in the probabilistic linkage (i.e., even deterministic links), SSN agreement status (defined as 7 or more matching digits) was used to measure Type I error. Type I error for probabilistic links was measured as the total number of probabilistic links with non-agreeing SSN divided by the total number of probabilistic links with SSN available on both the survey and administrative 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. Since a sizeable proportion of links were derived from the deterministic method, this had the effect of reducing the estimated Type I error by the proportion of probabilistically determined linkages among all linkages. For example, the Type I error rate was estimated for probabilistic links as 1.2%, but only 40% of all links were derived from probabilistic analysis. Thus, the estimated Type I error rate for the combined linkage process was (0.40*0.012) = 0.0048 or 0.48%.

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

4.2 Set Probability Cutoff

One goal of record linkage is to have the lowest errors possible. However, as more pairs were accepted, pairs that were less certain to be matches as links increase the Type I error and decrease Type II error (see Figure 3). And as less pairs were accepted, pairs that were more certain to be matches as links decrease the Type I error and increase Type II error. The optimal trade-off is between Type I error and Type II error was not known, and likely this depends on the type of analysis to be conducted with the linked data, but it is assumed that it is not far from optimality when the sum of Type I and Type II error is at a minimum. For this reason, Type I and Type II error are estimated at various probability cut points and the one that showed the lowest estimate of total error was selected. For this linkage, the probability cutoff was set to 0.92.

Figure 3: Error Level by Cutoff Value
(Schematic: not based on actual analysis)



4.3 Select Links Using Probability Threshold

The final step in the linkage algorithm was to determine links, which were pairs imputed to be matches. Links were pairs where the $Probvalid_{SSN_{Adj}}$ exceeded the set probability threshold (from Section 4.2). All pairs with an adjusted probability that fell below the set probability threshold were not linked.

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

4.4 Computed Error Rates of Selected Links

Final error rates were computed for selected links (described in Section 4.3). Table 8 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 2016 linked NHCS-CMS T-MSIS linkages. 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 NHCS record was a match to the CMS T-MSIS 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 (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 8. Algorithm Results for Total Selected Links

	Cutoff	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
2016 NHCS	0.92	2,163,586	467,103 (21.6%)	1,696,483 (78.4%)	0.04%	1.5%

Table 9 provides the total selected links, number of probabilistic and deterministic links, and the estimated Type I and II error rates for the selected links, by record type source for the 2016 NHCS. As shown in Table 9, UB-04 claims have higher estimated linkage error (both Type I and II) compared to the EHR records. Due to elevated levels of missing data in EHRs compared to the UB-04 claims records, the number of deterministic matches made by the algorithm for EHR Custom Extract (59.5%) is proportionally higher than UB-04 deterministic matches (18.3%). This resulted in a lower proportion of EHRs having CMS T-MSIS data extracted based on the probabilistic linkage. Additionally, CCD data were delivered without SSN information. This resulted in 100% of CCDs having CMS T-MSIS data extracted based on the probabilistic linkage and therefore the Type II linkage error rate was not calculated.

Table 9. Algorithm Results for Total Selected Links by 2016 NHCS Data Source

Data Source	Cutoff	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
UB-04 Claims	0.92	1,690,830	309,062 (18.3%)	1,381,768 (81.7%)	0.04%	2.0%
EHR Custom Extract	0.92	265,694	158,041 (59.5%)	107,653 (40.5%)	0.01%	0.4%
CCD	0.92	207,062	0 (0%)	207,062 (100%)	0.07%	*

*Unable to estimate Type II linkage error due to no SSN information on CCD records.

Appendix II: Assessment of 2015-2017 T-MSIS Identification Variables

1 Introduction

Prior to conducting a data linkage, an important first step is to assess the completeness of the variables used to link records from the two data sources at the person level. Because this was the first linkage of the NHCS data to the CMS Transformed Medicaid Statistical Information System (T-MSIS) administrative data files, an analysis of the completeness of T-MSIS identification variables was conducted. This information may be useful to the broader statistical community considering linking person-level data to T-MSIS. To enhance the utility of NCHS survey data collections, the standard NCHS data linkage algorithm attempts to use the following identification data elements collected from person-level survey data to link to health-related data sources: First and Last Name, Middle Initial, Date of Birth (month, day, year), Sex, Zip Code and State of Residence, and Social Security Number (all 9 digits or last 4 depending on availability).

Prior to undertaking the linkage of the 2016 NHCS to 2015-2017 T-MSIS data, NCHS conducted an assessment of the completeness of the T-MSIS identification variables to evaluate the missingness of the data necessary to conduct a person-level linkage. Because NCHS is linking national survey data to T-MSIS data, this assessment was conducted at the national level, rather than assessing individual states.

2 State level T-MSIS reporting

States began transitioning to reporting Medicaid data in the T-MSIS format beginning in 2014, and as of 2016, all 50 states, the District of Columbia (DC), and Puerto Rico were reporting T-MSIS data to CMS [51]. The U.S. Virgin Islands began reporting T-MSIS data in 2017. Since the linkage of the NHCS data to CMS Medicaid data includes the T-MSIS transition period, information on the number of states reporting in T-MSIS format for 2015-2017 is provided in Table 1. Additional information regarding which states submitted T-MSIS data in 2015 is available at TAF Research Identifiable File (RIF) Availability Chart (medicaid.gov) and is discussed further in Section 4.2.1.

3 Assessment of identification variables

Table 10 provides an assessment of identification variable completeness by variable type and year for all T-MSIS reporting states. Because there is an undetermined level of legitimate missingness for middle initial, its completeness was not assessed in this report. Overall, identifier variable completeness is above 87% for all reporting states combined, in all years. The completeness of all identification variable types improved from 2015 through 2017. By 2017, each of the identifier variables assessed were at least 93% complete.

Table 10. Percent of identifier variables that are available for use in T-MSIS record linkage, by year, and number of reporting states and territories*

Linkage Variable Name	2015	2016	2017
Social Security Number (SSN)	90.6	92.2	93.4
First Name	96.7	97.3	97.9
Last Name	97.5	97.6	98.2
Day of Birth	96.8	97.2	97.3
Month of Birth	96.8	97.2	97.3
Year of Birth	96.8	97.2	97.3
Sex	96.7	97.2	97.3
Zip Code	87.6	92.3	94.3
State of Residence	87.6	92.3	94.3
Number of States/Territories Submitting T-MSIS Data	31	52	53

*Identifier variable availability is defined as non-missing information on the Medicaid enrollee's enrollment record.

4 Conclusion

State reporting of identification variables in T-MSIS submissions has improved overall from 2015 through 2017. Given the overall completeness of the identifier variables at the national level, NCHS felt confident pursuing the linkage of its national survey data with T-MSIS. This assessment expands the public knowledge of the availability and completeness of commonly utilized linkage identification variables included in the T-MSIS Analytic Files.

References

[1] https://www.kff.org/wp-content/uploads/2013/01/8193.pdf (accessed June 10, 2022).

[2] https://www.medicaid.gov/medicaid/program-information/medicaid-and-chip-enrollment-data/report-highlights/index.html (accessed June 10, 2022).

[3] https://www.cms.gov/newsroom/fact-sheets/medicaid-facts-and-figures (accessed June 10, 2022).

[4] https://www.cms.gov/newsroom/press-releases/cms-office-actuary-releases-2019-national-health-expenditures (accessed June 10, 2022).

[5] https://www.ncsl.org/research/health/long-term-services-and-supports-faqs.aspx (accessed June 10, 2022).

[6] https://www.medicaid.gov/medicaid/benefits/behavioral-health-services/index.html (accessed June 10, 2022).

[7] https://www.kff.org/medicaid/issue-brief/medicaid-financing-the-basics/ (accessed June 10, 2022).

[8] https://www.macpac.gov/wp-content/uploads/2015/01/EXHIBIT-16.-Medicaid-Spending-by-State-Category-and-Source-of-Funds-FY-2020-millions.pdf (accessed June 10, 2022).

[9] https://www.cms.gov/files/document/nhe-projections-2019-2028-forecast-summary.pdf (accessed June 10, 2022).

[10] https://www.medicaid.gov/sites/default/files/2019-12/list-of-eligibility-groups.pdf (accessed June 10, 2022).

[11] http://childwelfaresparc.org/wp-content/uploads/2014/10/Medicaid-to-26-for-Former-Foster-Youth7.pdf (accessed June 10, 2022).

[12] https://www.medicaid.gov/medicaid/benefits/early-and-periodic-screening-diagnostic-and-treatment/index.html (accessed June 10, 2022).

[13] https://www.macpac.gov/characteristics-of-key-medicaid-managed-care-spas-and-waivers/ (accessed June 10, 2022).

[14] https://www.macpac.gov/medicaid-101/waivers/ (accessed June 10, 2022).

[15] https://www.kff.org/medicaid/issue-brief/medicaid-waiver-tracker-approved-and-pending-section-1115-waivers-by-state/ (accessed June 10, 2022).

[16] https://www.cms.gov/Outreach-and-Education/American-Indian-Alaska-Native/AIAN/LTSS-TA-Center/info/1915-c-waivers-by-state (accessed June 10, 2022).

[17] https://www.medicaid.gov/about-us/program-history/index.html (accessed June 10, 2022).

[18] https://www.medicaid.gov/chip/eligibility/index.html (accessed June 10, 2022).

[19] https://www.medicaid.gov/chip/state-program-information/index.html (accessed June 10, 2022).

[20] https://www.medicaid.gov/state-overviews/scorecard/annual-medicaid-chip-expenditures/index.html (accessed June 10, 2022).

[21] https://www.kff.org/medicaid/state-indicator/total-chip-spending/ (accessed June 10, 2022).

[22] https://www.ssa.gov/OP_Home/ssact/title21/2103.htm (accessed June 10, 2022).

[23] https://www.medicaid.gov/dq-atlas/downloads/supplemental/3011_Final_Action_Status.pdf (accessed June 10, 2022).

[24] https://www.macpac.gov/wp-content/uploads/2016/03/Medicaid-Inpatient-Hospital-Services-Fee-for-Service-Payment-Policy.pdf (accessed June 10, 2022).

[25] Swan, J. H., C. Harrington, L. A. Grant, "Reimbursement for Nursing Homes, 1978-86", Health Care Financing Review, Vol.9 No. 3, Spring 1988, p. 33-50. https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/HealthCareFinancciningReview/Downloads/CMS1192036dl.pdf (accessed June 10, 2022).

[26] https://www2.ccwdata.org/documents/10280/19002246/ccw-taf-rif-user-guide.pdf (accessed June 24, 2022)

[27] https://www.drugs.com/ndc.html (accessed June 10, 2022).

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

[29] https://www.nashp.org/medicaid-family-planning-demonstrations-design-issues-and-resources-states/ (accessed June 10, 2022).

[30] https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAX-Validation-Reports-Items/CMS1238400, (accessed June 10, 2022).

[31] https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAX-Validation-Reports-Items/CMS1238402 (accessed June 10, 2022).

[32] https://www.macpac.gov/topics/dually-eligible-beneficiaries/ (accessed June 10, 2022).

[33] https://www.markfarrah.com/mfa-briefs/managed-medicaid-enrollment-trends-and-market-insights/ (accessed June 10, 2022).

[34] Filtered Managed Care Enrollment Summary (medicaid.gov) (accessed June 10, 2022).

[35] https://www.medicaid.gov/medicaid/managed-care/enrollment-report/index.html (accessed June 10, 2022).

[36] https://www.medicaid.gov/dq-atlas/landing/topics/single/map?topic=g8m81&tafVersionId=24 (accessed June 10, 2022).

[37] https://www.medicaid.gov/dq-atlas/downloads/supplemental/7011_Shared_Medicaid_IDs_2016.pdf (accessed June 10, 2022).

[38] https://www.cms.gov/newsroom/fact-sheets/fact-sheet-medicaid-and-chip-t-msis-analytic-files-data-release (accessed June 10, 2022).

[39] https://www.medicaid.gov/status-of-t-msis-priority-items-1-12-of-july-2020/index.html (accessed June 10, 2022).

[40]https://www.medicaid.gov/status-of-t-msis-priority-items-1-23-of-july-2020/index.html (accessed June 10, 2022).

[41] https://www.medicaid.gov/medicaid/data-and-systems/macbis/tmsis/tmsis-blog/entry/54044 (accessed June 10, 2022).

[42] https://www.shadac.org/news/raceethnicity-data-cms-medicaid-t-msis-analytic-files-updated-february-2021-%E2%80%93-features-2018 (accessed June 10, 2022).

[43] https://resdac.org/ (accessed June 10, 2022).

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

[45] Christen, Peter. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications. Berlin Heidelberg: Springer-Verlag, 2012. http://www.springer.com/us/book/9783642311635 (accessed June 10, 2022).

[46] Michelson, Matthew, and Craig A. Knoblock. "Learning Blocking Schemes for Record Linkage." In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, 440–445. AAAI'06. Boston, Massachusetts: AAAI Press, 2006. https://pdfs.semanticscholar.org/18ee/d721845dd876c769c1fd2d967c04f3a6eeaa.pdf (accessed June 10, 2022).

[47] 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 (accessed June 10, 2022).

[48] 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.

[49] 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.

[50] 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 (accessed June 10, 2022).

[51] https://www.cms.gov/newsroom/fact-sheets/fact-sheet-medicaid-and-chip-t-msis-analytic-files-data-release (accessed June 10, 2022).