The Linkage of National Center for Health Statistics Survey Data to Centers for Medicare & Medicaid Services Transformed Medicaid Statistical Information System Claims Data (2014-2019): Matching Methodology and Analytic Considerations

Data Release: October 14, 2022

Document Version Date: September 27, 2024

Division of Analysis and Epidemiology

National Center for Health Statistics

Centers for Disease Control and Prevention

datalinkage@cdc.gov

Suggested Citation: National Center for Health Statistics. Division of Analysis and Epidemiology. *The Linkage of National Center for Health Statistics Survey Data to Centers for Medicare & Medicaid Services Transformed Medicaid Statistical Information System Claims Data (2014-2019): Matching Methodology and Analytic Considerations*, September 2024. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data-linkage/medicaid-methods.htm

Contents

List of Acronyms

AIDS, Acquired Immunodeficiency Syndrome

CCW, Chronic Conditions Warehouse

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

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

LSOA II, Second Longitudinal Study of Aging

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

NHANES, National Health and Nutrition Examination Survey

NHANES III, Third National Health and Nutrition Examination Survey

NHEFS, NHANES I Epidemiologic Follow-Up Study

NHHCS, National Home and Hospice Care Survey

NNHS, National Nursing Home Survey

OP, outpatient

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. As part of its ongoing efforts to fulfill this mission, NCHS conducts several population-based and establishment surveys that provide rich cross-sectional information on risk factors such as smoking, height and weight, health status, and socio-economic circumstances. Although the survey data collected provide information on a wide-range of health-related topics, they often lack information on longitudinal outcomes.

Through its Data Linkage Program, NCHS has been able to expand the analytic utility of the data collected from NCHS surveys by augmenting it with Medicaid and Children's Health Insurance Program (CHIP) 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 selected NCHS surveys to 2014-2019 CMS T-MSIS claims and enrollment data.** Linking NCHS survey data with information from T-MSIS 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.

2 Background on Linked Files

2.1 National Center for Health Statistics Survey Data

NCHS has recently linked data from following surveys to 2014-2019 CMS T-MSIS enrollment and claims data:

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

Additionally, NCHS previously linked data from the following NCHS surveys to 1999-2014 CMS Medicaid enrollment and claims data:

- 1994-2013 NHIS
- Second Longitudinal Study of Aging (LSOA II)
- 1999-2012 Continuous NHANES
- NHANES III
- NHANES I Epidemiologic Follow-Up Study (NHEFS)
- 2004 NNHS
- 2007 National Home and Hospice Care Survey (NHHCS)

More information regarding the Linked NCHS-CMS Medicaid linked data files with 1999-2014 Medicaid data can be found in the matching methodology and analytic considerations report that accompanied that data release.[1] Researchers should be aware that, although this previous linkage is still available for use in approved research proposals, the 1999–2014 linked Medicaid data is in the Medicaid Statistical Information System (MSIS) Medicaid Analytic Extract (MAX) format, which differs substantially from the current 2014-2019 T-MSIS data format.

A brief description of the NCHS surveys included in the CMS T-MSIS linkage follows.

2.1.1 NHIS

NHIS is a nationally representative, cross-sectional household interview survey that serves as an important source of information on the health of the civilian, noninstitutionalized population of the United States. It is a multistage sample survey with primary sampling units of counties or adjacent counties, secondary sampling units of clusters of houses, tertiary sampling units of households, and finally, persons within households. It has been conducted continuously since 1957 and the content of the survey is periodically updated. NHIS has been used as the sampling frame for other NCHS surveys focusing on specialized populations, including LSOA II. Prior to 2007, NHIS traditionally collected full 9-digit Social Security Numbers (SSN) from survey participants. However, in attempt to address respondents' increasing refusal to provide SSN and consent for linkage, in 2007 NHIS began to collect only the last 4 digits of SSN and added an explicit question about linkage for those who refused to provide SSN. The implications of this procedural change on data linkage activities are discussed later in this report. For detailed information on the NHIS's contents and methods, refer to the NHIS website, http://www.cdc.gov/nchs/nhis.htm (accessed September 19, 2022).

2.1.2 NHANES

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

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

For detailed information about the Continuous NHANES and NHANES III contents and methods, refer to the NHANES website, https://www.cdc.gov/nchs/nhanes/index.htm (accessed September 19, 2022).

2.1.3 NNHS

NNHS provides information on nursing homes from two perspectives- that of the provider of services and that of the recipient of care. Data for the surveys were obtained through personal interviews with facility administrators and designated staff who used administrative records to answer questions about the facilities, staff, services and programs, and medical records to answer questions about the residents. NNHS was first conducted in 1973-1974 and repeated in 1977, 1985, 1995, 1997, 1999, and 2004. Only the 2004 survey was included in this CMS T-MSIS linkage. For more information on the NNHS content and methods, refer to the NNHS website, http://www.cdc.gov/nchs/nnhs.htm (accessed September 19, 2022).

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 [2]. 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].

[1] For a legislative history of Medicaid and CHIP, see https://www.macpac.gov/reference-materials/federal-legislative-milestones-in-medicaid-and-chip/ (accessed September 19, 2022).
[2] For more information on Medicaid state plan requirements, see https://www.macpac.gov/subtopic/state-plan/ (accessed September 19, 2022).

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 [3]. 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 [4].

Medicaid accounts for almost one-sixth of national spending on personal health care [5]. Medicaid is the main payer of nursing home care and long-term care services overall [6]; it is also the largest source of public funding for mental health care [7]. Seniors and people with disabilities make up approximately 25% of all Medicaid enrollees but account for two-thirds of Medicaid benefit expenditures [8]. 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 [9], 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 [10]. 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 [11], 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 [12]
- Elderly and disabled individuals receiving Supplemental Security Income (SSI)
- 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]

[3] The U.S. population estimate for July 1, 2018 was 327.2 million individuals, 2018 National and State Population Estimates (census.gov) (accessed September 19, 2022).
[4] Income cut-offs are based on percentage of the federal poverty level (FPL) and vary by state.
[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.

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.

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
- 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

[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 September 19, 2022)
[8] For complete lists of mandatory and optional services, see https://www.medicaid.gov/medicaid/benefits/index.html (accessed September 19, 2022).

- 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
- 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 [13]. 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 [14]. 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 [15]. 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 [16].
- **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 [17]. 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 [18]. CHIP provides coverage for children who meet the following criteria [19]:

- 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 [20]:

- **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 [21]. The federal share of CHIP spending was 94.2 percent while state and local spending accounted for the remaining 5.8 percent [22].

CHIP covered services include [23]:

- 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

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 September 19, 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

There are five Medicaid/CHIP T-MSIS Analytic files (TAFs) available to analysts who have access to the Linked NCHS-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 4.3.5.

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

All five TAFs contain 6 years (2014-2019) 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 [24]. 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 September 19, 2022) to assist analysts in understanding how to analyze TAF research files, as well as TAF Technical Guidance documents for the DE file and for the claims files (accessed September 19, 2022).[25],[26]

Demographic and Eligibility 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. The DQ (Data Quality) Atlas (accessed September 19, 2022) produced by CMS includes information on the data quality of variables in the DE file.

Inpatient Hospital 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 OT file.

The IP File includes Diagnosis Related Group (DRG) codes which are used to reimburse inpatient hospital services in many states [27]. 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 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 [28]. 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 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. [29]

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 [30] 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 for additional information).

Other Services 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 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.

[10] Drug groupers are available from Wolters Kluwer Health, known as Medi-Span, and First Data Bank.
[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 September 19, 2022)

3 Linkage Methodology

3.1 Linkage Eligibility Determination

The linkage of NCHS survey participant data to CMS T-MSIS administrative records was conducted through an agreement between NCHS and CMS. Approval for the linkage was provided by the NCHS Research Ethics Review Board (ERB)[12] and the linkage was performed only for eligible NCHS survey participants. Only NCHS survey participants who have provided consent as well as the necessary personally identifiable information (PII), such as date of birth and full or partial SSN or Medicare Health Insurance Claim Number (HICN), are considered linkage-eligible. Linkage-eligibility refers to the potential ability to link data from an NCHS survey participant to administrative data. This is distinct from program eligibility, which defines whether a person meets the eligibility criteria for a specific government-administered or funded program. Due to variability of questions across NCHS surveys, changes to PII collection procedures by the surveys over time, and changes in who is asked specific questions, criteria for NCHS-CMS T-MSIS linkage eligibility vary by survey and year.

For many of the surveys initiated prior to and during 2007 (for NHIS) or 2008 (for NHANES), including 1994-2006 NHIS, 1999-2008 NHANES, NHANES III, and 2004 NNHS, a refusal by the survey participant to provide an SSN or HICN was considered an implicit refusal for data linkage. However, NCHS began to notice an increase in the refusal rate for providing SSN and HICN, particularly for NHIS, which reduced the number of survey participants eligible for linkage. [31] In an attempt to address declining linkage eligibility rates, NCHS began investigating new procedures for obtaining consent for linkage from survey participants. Research was also conducted to assess the accuracy of matching data from NHIS to the National Death Index (NDI) using partial SSN and other PII. [32] The research assessed algorithms using the last four and last six digits of SSN. The results were favorable and provided sufficient data to support changes in how NHIS collected SSN and HICN for linkage. [33] Beginning in 2007, NHIS started requesting only the last four digits of SSN and HICN (plus a 1 - 2 length alphanumeric code) instead of the complete number for both identifiers. In addition, a short introduction before asking for SSN was added and participants who refused to provide SSN or HICN were asked for their explicit permission to link to administrative records without SSN or HICN. Also, at this time, the NCHS ERB determined that for 2007 NHIS and all subsequent years, only primary respondents (sample adult and sample child) would be eligible for linkage to administrative records.

The informed consent procedures changed for the continuous NHANES as well. NHANES continued to collect full nine-digit SSN and complete HICN through the 2017-2018 survey cycle. However, beginning with the 2009-2010 NHANES, participants were explicitly asked for consent to be included in data linkage activities during the informed consent process prior to the interview. Only participants who provided an affirmative response to the linkage question were considered linkage eligible. In addition, starting in 2017-2018, survey participants who consented to linkage but who refused to provide their full nine-digit SSN and complete HICN were given the option to provide only the last four digits of either identification number.

[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.

3.2 Child Survey Participants

NCHS survey participants under 18 years of age at the time of the survey are considered linkage-eligible if the linkage eligibility criteria described above are met and consent is provided by their parent or guardian. However, the consent provided by the parent or guardian does not apply once the child survey participant becomes a legal adult, and there is no opportunity for NCHS to obtain consent to link the child participant's survey data to administrative data based on their adult experiences. As a result, in accordance with NCHS ERB guidance, NCHS only includes administrative data that were generated for program participation, claims and other events that occurred prior to calendar year in which the participant turned 18 years old on the linked data files provided to researchers. Researchers should consider the impact of this censoring as they develop their RDC proposal. For more information about how to identify linked child survey participants please see Section 4.3.3.

3.3 Overview of Linkage

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

Linkage-eligible NCHS survey participant records were linked to the CMS T-MSIS enrollment database using the following identifiers: SSN (9 digits or last 4 digits, depending on the survey and year of the survey), first name, last name, middle initial, month of birth, day of birth, year of birth, 5-digit ZIP code of residence, state of residence, and sex.

The NCHS survey participant records and the CMS 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.[34] 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.)
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 NCHS survey participant record that was linked, CMS extracted the T-MSIS claims information and sent the data to NCHS following secure data transfer procedures.

4 Analytic Considerations

This section summarizes some key analytic considerations for users of the linked NCHS-CMS T-MSIS claims records. It is not an exhaustive list of the analytic concerns that researchers may encounter while

using the linked NCHS-CMS 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 NCHS Survey Data

4.1.1 Sample Weights

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

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

A model-based calibration approach developed within the SUDAAN software package (Procedure WTADJUST or WTADJX) allows auxiliary information to be used to adjust the sample weights for nonresponse. This approach is recommended for adjusting sample weights for the linked files. Because inferences may depend on the approach used to develop weights, within SUDAAN's WTADJUST or using a different calibration approach, researchers should seek assistance from a statistician for guidance on their particular project. Other approaches or software can be used. NCHS continues to investigate alternate approaches for addressing issues related to missing data, including the use of multiple imputation techniques. More detailed information on adjusting sample weights for linkage eligibility using SUDAAN can be found in Appendix III of *Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare & Medicaid Services* [35] and in *Assessing Linkage Eligibility Bias in the National Health Interview Survey*.[36] To calculate the adjusted weights for linkage eligibility it is suggested that researchers use the TMSIS_MATCH_1419 variable from the T-MSIS Match File (see Section 4.3.3).

4.1.2 Survey Participant Identification Variables

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

4.2 Analytic Considerations for CMS T-MSIS Data Files

4.2.1 State Differences in T-MSIS Reporting

CMS began working with states in 2011 to transform the way states report Medicaid data to CMS from the existing national 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 for 19 transitioned states for calendar year 2014 and 30 transitioned states (including DC) in 2015.

Researchers should be aware that the Linked NCHS-CMS T-MSIS data files do not include state-submitted 2014 Medicaid data for the following 32 states: AR, AZ, CA, CT, GA, HI, IA, ID, IN, KY, LA, MA, MI, MN, MO, MS, NJ, NY, OH, OK, OR, PA, SC, SD, TN, TX, UT, VA, VT, WA, WV, and WY. The Linked NCHS-CMS T-MSIS data files do not contain 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 NCHS-CMS T-MSIS files include state-submitted T-MSIS data for all 50 states plus DC for 2016–2019.

Researchers who wish to use the Linked NCHS-CMS T-MSIS data for 2014 or 2015 should carefully consider the implications of this state-based missingness on their analytic assumptions and the interpretation of results.

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. As a reminder, NHIS and NHANES are designed to be nationally representative of the civilian noninstitutionalized population.

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.

To identify all enrollees in a given month, the TAF Technical Documentation recommends using a combination of variables CHIP_CD_01 through CHIP_CD_12 and ELGBLTY_GRP_CD_01 through ELGBLTY_GRP_CD_12. [25]

Researchers 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. A value of 1 for these variables identifies individuals who were not enrolled in either an M-CHIP or an S-CHIP program. A value of 2 for these variables identifies individuals who were enrolled in an M-CHIP program. A value of 3 identifies individuals who were enrolled in S-CHIP. A value of 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. Because some states do not populate CHIP_CD_mm for all enrollees, this variable may be missing. More detailed information on how to use other variables to obtain this missing information is provided in the DE TAF documentation. [25]

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. 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.[13] 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

[13] See TAF Methodology Brief #4151 https://requests.resdac.org/sites/resdac.umn.edu/files/4151_Scope_of_Benefits.pdf (accessed September 19, 2022) for details on identifying categories of benefit packages in the TAF.

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 [37]. 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 [38]. 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 NCHS-CMS 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 NCHS survey participants (see Section 5.4.1 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 1.

Table 1: 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 [39]. 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 [40]. 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 [41].

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 in the linked data.

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 NCHS-CMS T-MSIS data 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 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 and CHIP DQ Atlas (accessed September 19, 2022) includes an analysis of payment consistency between header and line-item claims records for the four claims file types [42]. 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[14]. 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

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

services provided to the mother, and services provided to the newborn may also be included in a single claim[15]. 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 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 [43].

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.

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 [44]
- The list expanded to a total of 23 items in 2019 [45]
- The list was again expanded to a total of 32 items in 2020 [46]

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

4.2.16.1 Assessment of T-MSIS Data Quality

CMS has produced a data quality assessment resource known as the Medicaid and CHIP DQ Atlas (accessed September 19, 2022) for each 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 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.

[15] See frequently asked questions #2437, #2463, and #3557 at https://www.cms.gov/files/document/frequently-asked-questions-9 (accessed September 19, 2022).

Subject matter topical areas discussed in the DQ Atlas 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 [47]. The Research Data Assistance Center (ResDAC) is also a valuable resource for information on T-MSIS data [48].

4.2.16.2 Reporting issues for Illinois Claims Data

There are special reporting issues that apply to T-MSIS claims data from the state of Illinois. Analysts 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 September 19, 2022).

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

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

NCHS survey participants may have multiple DE records. Most often, this is because a survey participant is linked to several years of T-MSIS data. However, a survey participant may be linked to multiple DE records within the same year. There are multiple explanations for this situation including survey participants 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 NCHS survey participants 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 NCHS-CMS Medicare data files. Service utilization records for services covered by Medicaid and not Medicare will be found in the Linked NCHS-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 NCHS-CMS T-MSIS data files. The NCHS surveys included in the Linked NCHS-CMS T-MSIS data files were previously linked to Medicare enrollment data for 2014–2018 and to Medicare FFS claims and Part D Prescription Drug Event data for 2016–2018.

Analysts interested in analyzing linked Medicare claims and prescription drug data for dually eligible individuals should request to use these Linked NCHS-CMS Medicare data files in their RDC proposal. For more information about the Linked NCHS-CMS Medicare data files, see Section 5.4.1.

4.3.3 T-MSIS Match Status File

The T-MSIS Match Status file can be used to identify which of the NCHS survey participants were eligible for linkage and linked to a T-MSIS DE record. This file contains one record for each NCHS survey participant and includes the variables TMSIS_MATCH_1419 and PROBVALID.

The variable TMSIS_MATCH_1419 should be used to determine linkage eligibility and match status (Section 3.1). NCHS survey participants with a TMSIS_MATCH_1419 value of 1, 2, or 3 were considered eligible for linkage to the T-MSIS DE records. This variable also indicates whether linkage-eligible NCHS survey participants linked to T-MSIS data. Both values 1 and 3 indicate the survey participant linked to at least one T-MSIS DE record from 2014–2019; however, value 3 indicates that linked data are only available for a child survey participant (Section 3.2) prior to the year in which they turned 18. A value of 2 indicates the survey participant was linkage eligible but did not link.

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 Linked NCHS-CMS T-MSIS data files, 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 5.3).

4.3.4 Temporal Alignment of Survey and Administrative Data

NCHS surveys have been linked to multiple years of T-MSIS administrative data. Depending on the survey year, T-MSIS data may be available for survey participants at the time of the survey, as well as before or after the survey period. Several factors may influence the alignment of the survey and administrative data, including residence state of the survey participant, program eligibility, and continuous program coverage. Users should be aware that linked NCHS survey participants may have linked Medicaid data for one or more years between 2014–2019, including the possibility of multiple non-continuous intervals.

4.3.5 Merging Within the Linked NCHS-CMS T-MSIS Data Files

Researchers should use the survey-specific survey participant identifier number (PUBLICID/SEQN/RESNUM) along with variables MSIS_SEQN[16], and FILE_YEAR4 to merge enrollment information from the DE base file to each of the TAF claims files. (See Appendix III: Merging Linked NCHS-CMS T-MSIS Files with NCHS Survey Data for more information on the survey participant identification number variable for each NCHS survey.)

The linked TAF claims files include separate files for claims header, line item, and occurrence[17]. To merge claim header, line, and occurrence information within a unique claim record, researchers should use the survey-specific survey participant identifier number along with 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 for an NHIS participant would merge data using the unique combination of PUBLICID, MSIS_SEQN, FILE_YEAR4, and NCHS_CLM_ID[18].

5 Access to the Restricted-Use Linked NCHS-CMS T-MSIS Claims Data Files

5.1 Obtaining Access

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 NCHS-CMS T-MSIS 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 September 19, 2022).

To create analytic files for use in the RDC, a researcher provides a file containing the variables from the public-use NCHS survey data to RDC for merging with the requested restricted variables from NCHS surveys and for use with the variables from the linked CMS T-MSIS data files. The full list of public-use variables, any restricted-use survey variables, and the exact variables from the linked CMS T-MSIS data files that the researcher will use also need to be specifically requested as part of a researcher's application to RDC. Staff in the RDC verify the full list of variables and check for potential disclosure risk.

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

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

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

5.2 Resources for Preparing an RDC Proposal

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 IP 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.

Each codebook also contains a link to the corresponding TAF codebook produced by the CMS Chronic Conditions Warehouse (CCW). Researchers are encouraged to review the CCW codebooks for more detailed descriptions, allowable values, and any updated information. The ResDAC website also contains data dictionary documentation for the TAFs.[48] Researchers are also strongly encouraged to review the TAF Technical Documentation documents for the DE file and the claims files when developing their research proposals.[25],[26]

5.3 Suggested Variables to Request in RDC Proposals

5.3.1 NCHS Survey Variables

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

- Survey participant identifier variables – Please refer to Appendix III: Merging Linked NCHS-CMS T-MSIS Files with NCHS Survey Data for guidance on identifying and constructing (if necessary) the appropriate identification variable.

- Sample weights and design variables—these variables are needed to account for the complex design of the NCHS surveys. The names of the weights and design variables differ depending on which NCHS survey is being used. These can be identified using the documentation for each NCHS survey. As discussed in Section 4.1.1, NCHS recommends adjusting the sample weights to account for linkage eligibility bias.

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

- Month and year of interview and/or NHANES examination— Many researchers will want to know the time elapsed between a given year (or even month) of the T-MSIS data and the survey events (NHIS interview; or NHANES interview or examination). NHANES is released in 2-year cycles. The exact year (and month) of a survey participant's interview and examination are not provided on public-use files. The variables that indicate the month and year of NHANES interview or examination must be requested specifically. For NHIS, the variables that indicate the interview month or quarter should be requested specifically.

5.3.2 Linked NCHS-CMS T-MSIS Variables

Although the complete list of variables used for specific analyses differs, the following linked NCHS-CMS T-MSIS variables should be considered for inclusion:

- Identification variables — Researchers should request both NCHS survey participant (e.g., PUBLCID, SEQN, RESNUM) and T-MSIS (e.g., MSIS_SEQN. FILE_YEAR4) variables in order to merge variables between the NCHS analytic files and the Linked NCHS-CMS T-MSIS claims data and when merging claims within the Linked NCHS-CMS T-MSIS claims datasets. Please see Sections 4.1.2 and 4.3.5 for more information on which identification numbers to request in your RDC proposal.

- Linkage eligibility and match status — To obtain information on NCHS survey participant eligibility for linkage, child participant status, and CMS T-MSIS match status, researchers should request variable TMSIS_MATCH_1419 from the NCHS-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).

- Medicaid and CHIP enrollment status — **Analysts proposing to analyze Linked NCHS-CMS T-MSIS claims data should request access to the DE TAF for the same calendar years as the Medicaid claims TAFs (IP, LT, RX, OT) in order to determine the correct study denominators for the linked Medicaid population.** To determine monthly enrollment, researchers will need to request the variables CHIP_CD_01 through CHIP_CD_12 and ELGBLTY_GRP_CD_01 through ELGBLTY_GRP_CD_12.[19] See Section 4.2.3 for more information regarding determining Medicaid program enrollment.

- Variables to identify enrollee subgroups of interest — Researchers are encouraged to review Section 4.2 for additional variables that may be needed to identify T-MSIS enrollee subgroups of interest, such as CHIP enrollees, restricted benefit enrollees, or dually eligible individuals.

- T-MSIS Submitting State — 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 Linked NCHS-CMS T-MSIS 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.

5.4 Additional Related Data Sources

5.4.1 Linked NCHS-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 Linked NCHS-CMS Medicare data (accessed September 24, 2024) for enrollment and fee-for-service claims data from 2014–2018and Medicare Advantage encounter data from 2016–2018. 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

[19] Data users familiar with Medicaid data may notice the Maintenance Assistance Status and Basis of Eligibility group variables (MASBOE_CD_xx) are on the DE file. However, these variables have been retired, and researchers are advised to use the Eligibility Group Code variables (ELGBLTY_GRP_CD_xx) for information regarding eligibility.

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 NCHS-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 Linked NCHS-CMS Medicare data files include health care claims and encounters, prescription drug events, and patient assessment data for linked Medicare beneficiaries for select years. To integrate the Linked NCHS-CMS Medicare linked data files into the Linked NCHS-CMS T-MSIS data files, joins are made on the survey-specific survey participant identification number (see Section 4.1.2).

More information about the Linked NCHS-CMS Medicare data files can be found at: https://www.cdc.gov/nchs/data-linkage/medicare.htm (accessed September 19, 2022).

5.4.2 Linked NCHS-NDI Mortality Files

Analysts interested in studying mortality among NCHS survey participants enrolled in Medicaid are encouraged to use the linked mortality data available in the NCHS-NDI Mortality files rather than the mortality data available in the Linked NCHS-CMS T-MSIS files. The linked NCHS-NDI Mortality files (accessed September 19, 2022) include information on deaths through December 31, 2019 identified for NCHS survey participants through linkage with the National Death Index and are not limited to deaths among the Medicaid enrolled population. In addition, in the NCHS-NDI linked data cause of death is available for survey participants who died. The linked mortality files include survey participant identification number, date of birth, date of death, and cause of death information for linked decedents. To integrate the linked NCHS-NDI linked data files into the Linked NCHS-CMS T-MSIS data files, joins are made on the survey-specific survey participant identification number (see Section 4.1.2).

More information about the linked NCHS-NDI Mortality data files can be found at: https://www.cdc.gov/nchs/data-linkage/mortality.htm (accessed September 19, 2022).

5.4.3 Linked NCHS-Housing and Urban Development (HUD) Administrative Data Files

Researchers interested in outcomes related to housing insecurity may also request variables from the Linked NCHS–HUD Administrative data files (accessed September 19, 2022) if federal housing assistance is a variable/outcome of interest. 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 NCHS–HUD administrative data files into the Linked NCHS-CMS T-MSIS data files, joins are made on the survey-specific survey participant identification number (see Section 4.1.2).

5.4.4 Linked NCHS-Department of Veterans Affairs (VA) Data Files

Researchers interested in outcomes related to Veterans may also request variables from the Linked NCHS-VA data files (accessed September 19, 2022). The Linked NCHS-VA data files include information on a wide range of health-related topics for Veterans, including Veteran status and utilization of VA benefit programs. To integrate the Linked NCHS-VA data files into the Linked NCHS-CMS T-MSIS data files, joins are made on the survey-specific survey participant identification number (see Section 4.1.2).

Appendix I: Detailed Description of Linkage Methodology

1 NCHS and CMS T-MSIS Linkage Submission Files

Prior to the linkage of the NCHS surveys 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 NCHS survey 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 survey participant (NCHS surveys) or enrollee (CMS T-MSIS)):

- SSN (validated)[20]
- 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 survey participants 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 I below provides two examples of how multiple part name information was used to generate alternate records, using hypothetical data. For survey participant A, the first name was used to generate multiple records, and for survey participant B, the last name was used.

[20] 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. For some surveys and survey years, only the last 4-digits (SSN4) were collected from survey participants. For SSN4 the last 4-digits cannot be 0’s (i.e. xxx-xx-0000).

Table I. Example of Alternate Record Generation using Name Fields

Survey participant 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 NCHS survey participant records and for CMS T-MSIS enrollment records, separately. During this process, multiple submission records were created for each survey participant/enrollee to show all combinations of the recorded values for these fields. That is, if a survey participant/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 survey participant/enrollee (see Table II for example).

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

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

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

2 Deterministic Linkage Using Unique Identifiers

The deterministic linkage, which was the next step in the linkage process, used only the NCHS 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 (SSN9) agreement and then for records where the last 4-digits of the SSN (SSN4) 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 (1st pass using SSN9) or greater than 2/3 (2nd pass using SSN4), the linked pair was retained as a deterministic match. In addition to the 2/3's agreement ratio, linked pairs in the 2nd pass were required to have at least 5 non-missing PII variables in agreement to be deemed a deterministic match. Of note, NCHS survey participants were excluded from the second pass (i.e., using SSN4) 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.[34] 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)." [49] 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 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." [50], [51]

The machine learning algorithm learned 14 blocking passes to be used in the blocking scheme. Table III provides the PII variables that were assigned to each of the blocking passes and the PII variables that were used to score the potential links in each of the blocking passes. Note, the variables listed in the scoring key are all PII variables not used as a blocking variable 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 III. 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 Appendix I Section 3.3), 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 III). Within the blocking pass, pairs with agreeing SSN - were used to calculate the M-probabilities, as these are assumed to represent the same individual. For records with a SSN9, agreeing SSN was defined as 8 or more digits being the same. For records with a SSN4, we required all 4 digits to be in agreement and at least 5 PII variables agreeing, totaling more than 2/3's agreement of all non-missing PII variables. For example, if we have a record with 6 non-missing PII variables and 5 agree, this would be kept for M-probability estimation. However, if we have a record with all 8 non-missing and 5 agree, this would not be used for M-probability estimation since it does not meet the 2/3 agreement requirement (i.e., 5/8=0.625). 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 survey participant ID and CMS T-MSIS ID (see Tables IV and V for example of record summarization).

Table IV is an example of how the agreement flags for each of the scoring variables in Blocking pass 3 are created. A value of 1 means the information in the variable is exactly matching, while a 0 means they are not. Table V then represents how the multiple submission records in table 5 are summarized into one record for each survey participant and administrative ID. If any of the identifiers agree across multiple records, they are flagged as agree (i.e., set to 1). The summarized records in table V are then used to estimate the M-probabilities for each of the specific scoring variables. For example, among qualifying pairs in blocking pass 3, 99.4% (M-probability Day Birth=0.994) agree on day of birth and 94.5% (M-probability ZIP=0.945) agreed on ZIP code of residence.

Table IV. Example of Agreement Flags Using Blocking Pass 3 as an Example

Survey Participant ID	CMS T-MSIS ID	Day of birth	Month of birth	Year of birth	ZIP Code	Middle Initial
1	1	1	0	1	0	.
1	1	.	1	1	0	0
1	1	1	0	1	0	0
2	2	1	0	1	0	0
3	789	1	1	.	0	1
3	789	0	1	0	1	1
3	789	.	1	0	1	.
3	789	0	0	1	1	1
3	322	1	0	1	1	1

NOTES: Data have been fabricated for the purposes of this example. PII, personally identifiable information
[1]Agreement status of 1 = match, 0 = non-match, and . = missing values

Table V. Example Showing Summarization of Blocked Records for M-Probability Estimation, Based on Records in Table IV

Survey Participant ID	CMS T-MSIS ID	Day of birth	Month of birth	Year of birth	ZIP Code	Middle Initial
1	1	1	1	1	0	0
2	2	1	0	1	0	0
3	789	1	1	1	1	1
3	322	1	0	1	1	1

NOTES: Data have been fabricated for the purposes of this example. PII, personally identifiable information
[1]Agreement status of 1 = match, 0 = non-match, . = missing values

Several additional comparison measures were created for first and last name and ZIP code 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 for records with SSN9 values and if any digits were not in agreement for records with SSN4 values). 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 NCHS 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 NCHS 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. [52], [53], [54]

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:

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

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

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 survey participant 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 used in the first iteration was set to half of the pairs in the blocking pass (i.e., all pairs generated by the blocking pass specification). The number of non-matches was computed by subtracting the estimated number of matches from the number of pairs (regardless of how likely they are to be matches) in the blocking pass.

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

Note that in the first iteration, it was assumed that the number of matches (within blocking pass B) were equal to the number of non-matches (within blocking pass B), resulting in $Adj_B = 0$. If, however, in a later iteration, the number of matches was estimated to be 20,000 and the number of pairs is 1,000,000, then

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

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

   $Odds_{1,B} = 2^{(8.4+\ -5.61)} \approx 6.9$

   if for Pair 2 of blocking pass B, the pair-weight is -2.5, then

   $Odds_{2,B} = 2^{(-2.5+\ -5.61)} \approx 0.0036$

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

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

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

   Continuing with the example…

   For Pair 1 in blocking pass B,

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

   For Pair 2 in blocking pass 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.

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

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

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

This process was repeated until convergence was reached in the number of matches being estimated. Once convergence was achieved, the final probabilities were estimated based on the last value of the number of matches (within blocking pass B) to be estimated. These estimated probabilities were then used to select the final matches, as described below in Section 4.

3.4 Adjustment for SSN Agreement

Up to this point, every pair generated through the probabilistic routine was assigned a value that estimates its probability of being a match. However, this estimate did not take SSN agreement into account. This was conducted as a separate step because for the other comparison variables, M- and U-probabilities were estimated based on probable matches or non matches that were determined based on SSN agreement and clearly this was infeasible for SSN itself.[21]

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 NCHS survey record and CMS T-MSIS record, the estimated probability was adjusted based on the last four digits of the SSN.[22]

When the last four digits of SSN[23] agreed (i.e., are exactly the same):

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

When the last four digits of SSN did not agree:

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

No adjustment was made for pairs that did not have an SSN on either the NCHS or CMS T-MSIS record. So, for these pairs:

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

[21] 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 NCHS and CMS T-MSIS record. The U-probabilities are estimated as the random chance that a 4-digit SSN value will agree, or simply $\frac{1}{9{,}999} \approx 0.0001$.

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

[23] 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.

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

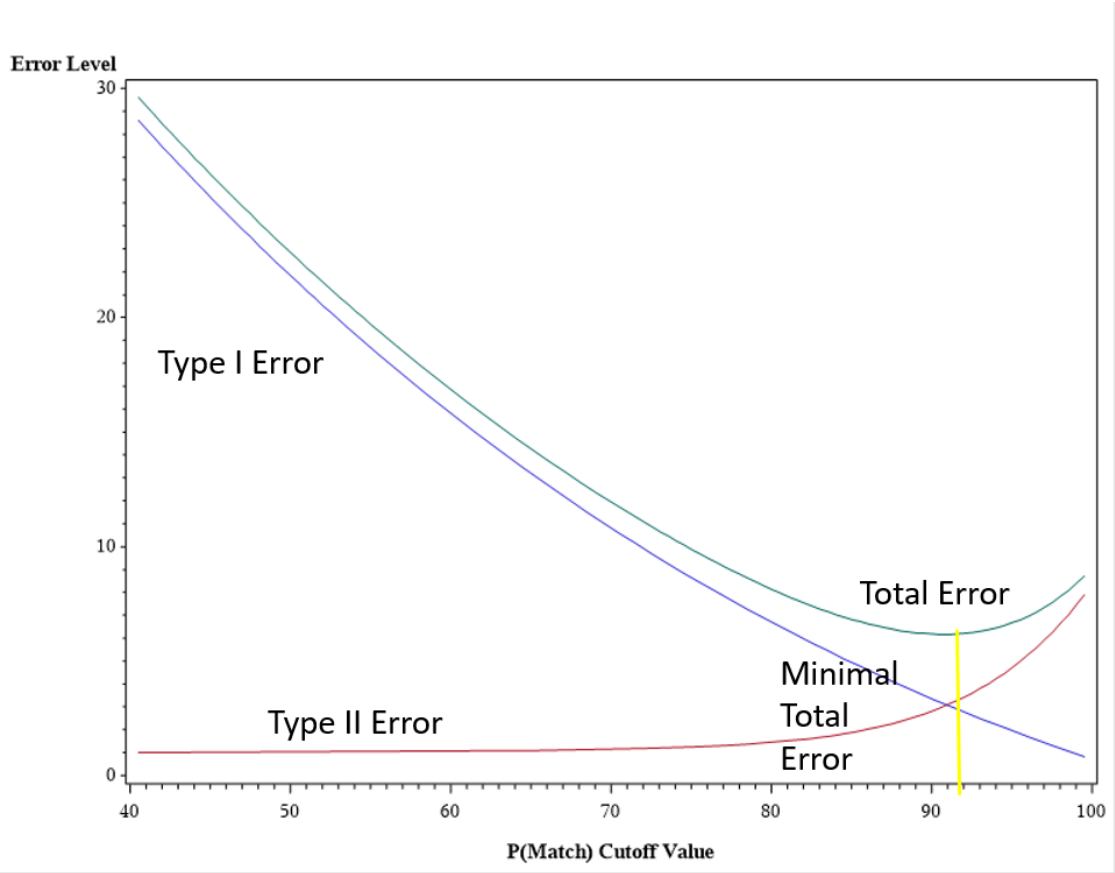
Because all records were included in the probabilistic linkage (i.e., even deterministic links), SSN agreement status (defined as 7 or more matching digits for records with SSN9 and all 4 digits for records with SSN4) was used to measure Type I error. Type I error for probabilistic links was measured 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, if the Type I error rate were estimated for probabilistic links as 1.2%, but only 40% of all links were derived from probabilistic analysis, then the estimated Type I error rate for the combined linkage process would be (0.40*0.012) = 0.0048 or 0.48%.

To measure Type II error, 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 would be 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 I). 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 I: 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 survey participant (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 VI provides the total number of selected links, the number of total links identified through deterministic and probabilistic methods, and the Type I and Type II error rates for the NCHS-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 NCHS 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 VI. Algorithm Results for Total Selected Links

	Cutoff	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
NCHS Surveys	0.92	277,801	158,089 (56.9%)	119,712 (43.1%)	0.06%	1.8%

Appendix II: Assessment of 2014-2019 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 NCHS national population survey participant 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 NCHS national population surveys to 2014–2019 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 nationally representative 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 [55]. The U.S. Virgin Islands began reporting T-MSIS data in 2017. Since the linkage of the NCHS survey data to CMS Medicaid data includes the T-MSIS transition period, information on the number of states reporting in T-MSIS format by year is provided in Table 8. Additional information regarding which states submitted T-MSIS data in 2014 and 2015 is available at TAF Research Identifiable File (RIF) Availability Chart (accessed September 19, 2022) and is discussed further in Section 4.2.1.

3 Assessment of identification variables

Table VII 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 across all years after 2015. The slight decrease in identification variable completeness in 2015 can be attributed to the increase in new states reporting T-MSIS data for the first time in that year. By 2019, each of the identifier variables assessed were at least 95% complete, except for SSN (93.4%).

Table VII. 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	2014	2015	2016	2017	2018	2019
Social Security Number (SSN)	96.3	90.6	92.2	93.4	93.8	93.4
First Name	98.0	96.7	97.3	97.9	98.3	98.6
Last Name	99.4	97.5	97.6	98.2	98.6	98.6
Day of Birth	98.2	96.8	97.2	97.3	97.7	98.1
Month of Birth	98.2	96.8	97.2	97.3	97.7	98.1
Year of Birth	98.2	96.8	97.2	97.3	97.7	98.1
Sex	98.2	96.7	97.2	97.3	97.7	98.0
Zip Code	92.1	87.6	92.3	94.3	95.1	95.8
State of Residence	92.1	87.6	92.3	94.3	95.1	95.8
Number of States/Territories Submitting T-MSIS Data	19	31	52	53	53	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 2014 through 2019. 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.

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

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

Note: The Linked NCHS-CMS T-MSIS data files are only available for research use through the NCHS restricted access data center (RDC). Approved RDC researchers may choose to provide their own analytic files created from public-use survey files to the RDC. Therefore, it is important for researchers to include survey specific Public Identification number on any analytic files sent to the RDC. The RDC will merge data (using PUBLICID, SEQN or RESNUM) from the linked CMS T-MSIS files to the analyst's file. The merged file will be held at the RDC and made available for analysis.

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

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

1.1 NHIS, 1994

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

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

SAS example:

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

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

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

1.2 NHIS, 1995-1996

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

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

SAS example:

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

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

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

1.3 NHIS, 1997-2003

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

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

SAS example:

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

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

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

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

1.4 NHIS, 2004

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

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

SAS example:

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

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

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

1.5 NHIS, 2005-2018

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

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

SAS example:

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

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

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

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

Item	Length	Description
SEQN	6	Participant identification number

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

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

Item	Length	Description
SEQN	5	Participant identification number

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

4 National Nursing Home Survey (NNHS), 2004

Item	Length	Description
RESNUM	6	Resident Record (Case) Number

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

Appendix IV: Concordance Between Self-Report of Medicaid Enrollment in the National Health Interview Survey, 2016–2018, and Medicaid Administrative Records

1 Introduction

Previous studies have assessed the accuracy of Medicaid reporting in surveys and have found a persistent, well-established undercount of Medicaid enrollment in national surveys in the range of 10%-30%.[56 57 58 59 60] In addition, a previous analysis by NCHS assessed the concordance of survey reported Medicaid enrollment in the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES) and Medicaid administrative data, the Medicaid Analytic eXtract files (MAX), among children and found an undercount of 11% in NHIS and 12% in NHANES,[61] in line with other estimates.

Understanding the potential for survey misreporting of Medicaid enrollment remains important for users of survey data. This analysis examines the agreement between survey report of Medicaid/Children's Health Insurance Program (CHIP) enrollment and administrative Medicaid/CHIP enrollment in linked NHIS and Transformed Medicaid Statistical Information System (T-MSIS) data. One motivation for updating the analysis is that T-MSIS is the newest source of Medicaid administrative data, which includes more detailed enrollment information and contains a broader set of beneficiaries (including separate CHIP and adult expansion populations) compared to the previous MAX files. Additionally, enhancements to improve linkage accuracy have been made to the linkage methodology used by the NCHS Data Linkage Program since the previous MAX linkage was conducted. To our knowledge, no previous studies have assessed the concordance of survey report of Medicaid/CHIP enrollment using the new T-MSIS reporting system.

2 Linked NHIS and CMS T-MSIS Data

2.1 Medicaid/CHIP Coverage Information

NHIS years 2016-2018 linked to Medicaid/CHIP enrollment data from T-MSIS for the same period were used for this analysis. Descriptions of the NHIS sample design are available in Section 2.1.1 and descriptions of Medicaid and T-MSIS are available in Section 2.2.

In NHIS, Medicaid enrollment information is based on the family respondent's answers to health care coverage question for themselves as well as each family member. The question reads, "What kind of health insurance or health care coverage {do/does} {person} have? INCLUDE those that pay for only one type of service (nursing home care, accidents, or dental care). EXCLUDE private plans that only provide extra cash while hospitalized." Any mention of health insurance is recorded, including Medicaid or CHIP. A respondent may list more than one source of coverage. If no source of coverage is specified for an individual age 64 and younger, the following probe question is administered that provides the specific state name of the Medicaid program: "There is a program called Medicaid that pays for health care for persons in need. In this State it is also called {*fill State name}. {Are you/Is ALIAS} covered by Medicaid?".[62] After the conclusion of the interview, NHIS applies an adjudication process that addresses conflicting information (e.g., if a respondent stated they have a private comprehensive plan and provided the name as 'Medicaid'). This editing process creates recoded variables classifying respondent's health insurance coverage, including variables indicating Medicaid or CHIP coverage.[24]

[24] NHIS does not distinguish between separate CHIP, Medicaid expansion CHIP, or combination CHIP.

Documentation from NHIS strongly recommends the use of these recoded variables for estimates of health care coverage.[63] Because children often cycle between Medicaid and CHIP coverage as their family circumstances and state coverage policies change, for the analysis of children the variables indicating Medicaid and CHIP coverage were combined into a single variable. If these recoded variables from the NHIS PERSON file indicated any Medicaid or CHIP coverage, the survey participant was categorized as having Medicaid/CHIP.

In T-MSIS, the Demographic and Eligibility (DE) analytic file includes eligibility and enrollment information for each beneficiary who was enrolled in Medicaid and/or CHIP for at least one day in a given calendar year. CMS recommends[25] using a combination of the monthly CHIP code variable and the monthly eligibility group variable to identify Medicaid beneficiaries, because not all states populate the CHIP code. The monthly restricted benefits code can be used to determine the level of benefits a beneficiary receives (full-scope, comprehensive, or restricted/limited).[64] Following the approach of other studies,[57] we defined Medicaid or CHIP health insurance coverage as including only full-scope or comprehensive plans, and excluded beneficiaries with restricted benefits only, as a survey respondent may not consider such limited plans to be sources of health insurance coverage.[65] Beneficiaries were classified as having a full-scope or comprehensive plan when the monthly restricted benefits code (RSTRCT_BNFTS_CD_mm) had a value of 1, 7, A, D, 4, or 5 in the same month as their NHIS interview.[25]

Because this analysis combined Medicaid and CHIP beneficiaries into one category, we classified an individual as having a T-MSIS report of Medicaid/CHIP enrollment if the CHIP code or the eligibility group variable indicated any Medicaid or CHIP enrollment[66] during the month of the NHIS interview. CHIP enrollment primarily applies to children, however there are limited circumstances where an adult may be eligible for CHIP, including those aged 18 years and pregnant women in some states.[19]

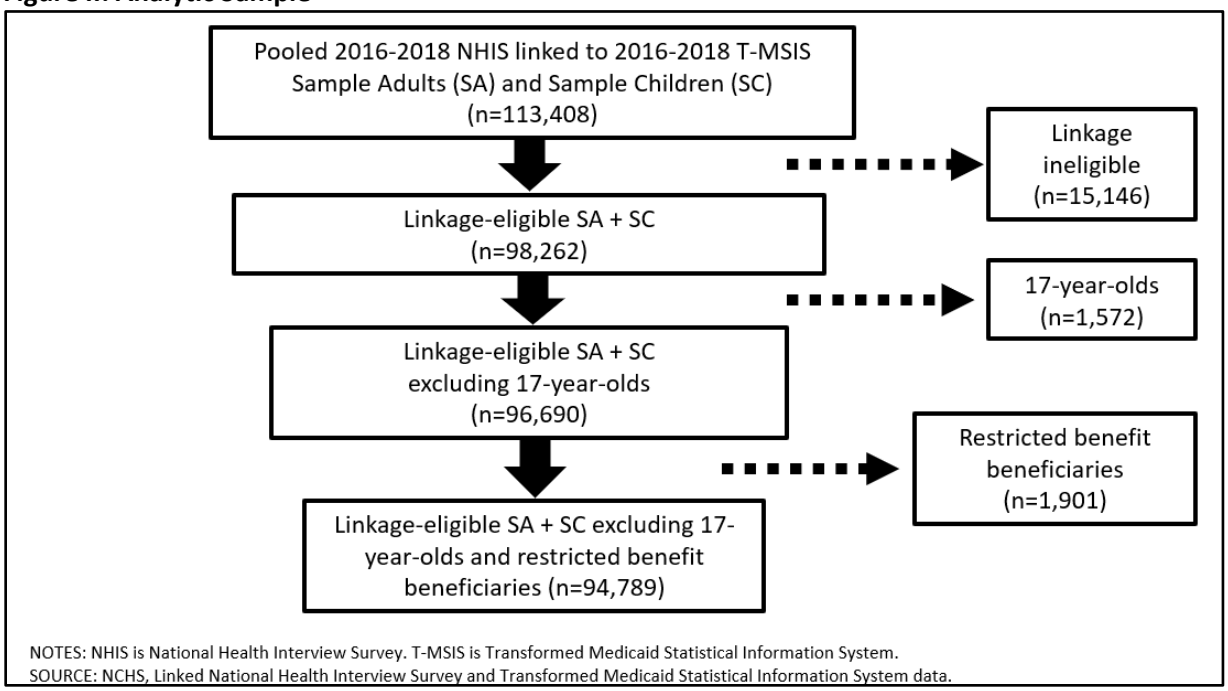
2.2 Analytic Sample

The analysis was limited to linkage-eligible sample adult and sample child participants from the 2016-2018 NHIS (see Figure II). Please see Section 3.1 for more information on linkage eligibility. Of the 113,408 sample adults and sample children interviewed in the 2016-2018 NHIS, 15,146 survey participants were excluded due to ineligibility for linkage. Sample children who were age 17 (n=1,572) at the time of the NHIS interview were also excluded from the analysis to account for potentially censored enrollment information due to the NCHS ERB consent protocol for children (see Section 3.2).[26] Finally, 1,901 survey participants who linked but were found to have restricted benefits at the time of their NHIS interview were excluded, leaving 94,789 participants in the analytic sample.

[25] The values for codes 4 and 5 vary by state; code 4 indicates full-scope coverage for all but three states, and the meaning of code 5 has varied over time and may indicate restricted benefits. Less than 0.5% of survey participants who linked to T-MSIS had a RSTRCTD_BNFTS_CD of 4 or 5 during the month of their interview.

[26] This protocol specified consent to link data provided by a parent or guardian on behalf of a child does not apply once the child becomes a legal adult (age 18). Therefore, on the NHIS-T-MSIS linked data files, Medicaid administrative data is only included for events that occurred prior to the calendar year in which the child survey participant turned 18. T-MSIS enrollment information as of the interview month and year would be censored for any 17-year-old survey participant who was interviewed during the calendar year in which they would turn 18. To avoid apparent discordance that was a result of this censoring, all 17-year-olds were removed from this analysis.

Figure II: Analytic Sample



3 Analytic Methods

We defined concordance as agreement between T-MSIS and NHIS in the same month and year of survey interview. Survey participants in the analytic sample were classified into one of four groups based on the agreement between T-MSIS and NHIS reporting of Medicaid/CHIP coverage:

1. Concordant (Y/Y): Yes–T-MSIS record of Medicaid/CHIP coverage and Yes–NHIS report of Medicaid/CHIP coverage.

2. Discordant (Y/N): Yes–T-MSIS record of Medicaid/CHIP coverage and No-NHIS report of Medicaid/CHIP coverage.

3. Discordant (N/Y): No–T-MSIS record of Medicaid/CHIP coverage and Yes–NHIS report of Medicaid/CHIP coverage.

4. Concordant (N/N): No–T-MSIS record of Medicaid/CHIP coverage and No-NHIS report of Medicaid/CHIP coverage.

Our analysis assessed agreement between T-MSIS and NHIS report of Medicaid/CHIP coverage overall (Table VIII) and by age groups: 0–16 years, 18–64, and 65 and older (Tables IX-XI). Cohen's kappa statistic was used to measure agreement between the T-MSIS administrative record and the survey report, and survey weights were not applied. A kappa statistic with a value of 0.60–0.79 is considered to have moderate agreement, 0.80–0.90 is considered strong, and above 0.90 is considered nearly perfect agreement [67]. Additionally, under the assumption that the T-MSIS report is the gold standard, the following statistics [68] were used to summarize the relationship between T-MSIS and NHIS reporting:

• Overall agreement: The proportion of participants who had concordant reports of Medicaid/CHIP coverage (Yes-T-MSIS and Yes-NHIS or No-T-MSIS and No-NHIS) among all participants.

• Sensitivity: The proportion of participants with a T-MSIS record of Medicaid/CHIP coverage who also had a survey report of Medicaid/CHIP coverage.

• Specificity: The proportion of participants with no T-MSIS record of Medicaid/CHIP coverage who also had no survey report of Medicaid/CHIP coverage.

• Positive predictive value: The proportion of participants with a survey report of Medicaid/CHIP coverage who also had a T-MSIS record of Medicaid/CHIP coverage.

• Negative predictive value: The proportion of participants with no survey report of Medicaid/CHIP coverage who also did not have a T-MSIS record of Medicaid/CHIP coverage.

All counts in the tables are unweighted, and all summary statistics are based on unweighted counts.

4 Results

Overall, among the 94,789 linkage-eligible sample adult and sample child participants, excluding 17-year-olds and beneficiaries with restricted benefits only, data for 14,177 (15.0%) participants were concordant Yes–T-MSIS and Yes–NHIS; 4,216 (4.4%) were discordant Yes–T-MSIS and No–NHIS; 2,187 (2.3%) were discordant No–T-MSIS and Yes–NHIS; and 74,209 (78.3%) were concordant No–T-MSIS and No–NHIS (Table VIII). The overall agreement was 93.2% (Yes–T-MSIS and Yes–NHIS and No–T-MSIS and No–NHIS) and the kappa statistic was 0.77 (standard error [SE] = .003). The sensitivity was 77.1%, and the specificity was 97.1%. The positive and negative predictive values were 86.6% and 94.6%, respectively.

This pattern of agreement persists by age group, though there are some key differences. Among the 22,228 NHIS sample children ages 0-16 included in the analysis, data for 7,033 (31.6%) participants were concordant Yes–T-MSIS and Yes–NHIS, 1,645 (7.4%) were discordant Yes–T-MSIS and No–NHIS, 929 (4.2%) were discordant No–T-MSIS and Yes–NHIS and 12,621 (56.8%) were concordant No–T-MSIS and No–NHIS (Table IX). The overall agreement was 88.4% and the kappa statistic was 0.75 (SE=.005), while the sensitivity was 81.0% and the specificity was 93.1%. The positive and negative predictive values were 88.3% and 88.5%, respectively.

For the 53,161 sample adults ages 18-64 included in the analysis, 6,004 (11.3%) participants were concordant Yes–T-MSIS and Yes–NHIS, 2,046 (3.8%) were discordant Yes–T-MSIS and No–NHIS, 949 (1.8%) were discordant No–T-MSIS and Yes–NHIS, and 44,162 (83.1%) were concordant No–T-MSIS and No–NHIS (Table X). The overall agreement was 94.4% and the kappa statistic was 0.77 (SE=.004). Sensitivity and specificity were 74.6% and 97.9%, respectively. The positive predictive value was 86.4% and the negative predictive value was 95.6%.

Lastly, for the 19,400 sample adults aged 65 and older included in the analysis, 1,140 (5.9%) participants were concordant Yes–T-MSIS and Yes–NHIS, 525 (2.7%) were discordant Yes–T-MSIS and No–NHIS, 309 (1.6%) were discordant No–T-MSIS and Yes–NHIS, and 17,426 (89.8%) were concordant No–T-MSIS and No–NHIS (Table XI). The overall agreement was highest of all age groups at 95.7%, but the kappa statistic was 0.71 (SE=.01) and the sensitivity was the lowest of all age groups at 68.5%. The specificity was 98.3 %, the positive predictive value was 78.7% and the negative predictive value was 97.1%.

5 Discussion

The agreement between T-MSIS and NHIS report of Medicaid or CHIP coverage was moderate based on the results of this analysis. While the overall agreement was 93.2%, the sensitivity of NHIS report of Medicaid or CHIP coverage compared to T-MSIS was 77.1%. This indicates that of survey participants who linked to a T-MSIS record indicating Medicaid or CHIP enrollment, 22.9% did not report this as their

coverage in NHIS. When limiting the population to survey participants ages 0–16 years, the overall agreement was 88.4% but the sensitivity was 81.0%, suggesting 19% did not correctly identify Medicaid or CHIP coverage in their NHIS responses to insurance coverage. For survey participants ages 18–64, overall agreement was 94.4% and sensitivity was 74.6%, indicating that approximately a quarter of adults in this age range did not correctly indicate Medicaid coverage in NHIS. Finally, for participants age 65 and older, the overall agreement was 95.7% but the sensitivity was 68.5%. Therefore, 31.5% of this age group were not correctly identified as Medicaid beneficiaries in NHIS.

Concordance was defined as agreement between T-MSIS and NHIS in the exact same month and year of the survey interview. Exploratory analyses that broadened the timespan for concordance did not meaningfully affect the results (data not shown).

This analysis has several limitations. First, it is important to note that this analysis was based on unweighted counts and so should not be used to infer population totals. Also, linkage error can occur. However, linkage error is unlikely to materially account for the discordant coverage categories due to low estimated type I (0.06%) and type II (1.8%) error rates reported in the NHIS-CMS T-MSIS data linkage.

It is possible that the decision to exclude beneficiaries with restricted benefits in T-MSIS and define Medicaid or CHIP health insurance coverage as including only full-scope or comprehensive plans increased the size of discordant Yes-NHIS No-T-MSIS group. However, a sensitivity analysis (data not shown) suggested this effect was minimal, and this analytic decision is consistent with other studies.

This analysis was conducted at the national level and does not address potential variation at the state level. Additionally, it should be noted that several states expanded Medicaid coverage during the study period.[69] It is possible the Medicaid expansion group differs from other Medicaid beneficiaries both in terms of self-report of Medicaid coverage as well as being captured in the administrative data. We used the standard NHIS recode variables MEDICAID and SCHIP, which do not include respondents who indicated they are covered by a ‘state-sponsored health plan’ other than Medicaid or CHIP. Lastly, the study population is not representative of all Medicaid beneficiaries. Institutionalized beneficiaries are not included in the NHIS sampling frame, and therefore are not included in this analysis.

Beginning in 2019, NHIS implemented a redesign that removed the family questionnaire entirely and moved the health insurance assessment questions to the Sample Adult Core and the Sample Child Core interviews. In the years of this analysis, the family respondent may or may not be the selected sample adult or be conferring with the selected sample adult when completing the health insurance assessment. While the wording of the questions was not changed in the redesign, it is possible that concordance may be higher in more recent years of NHIS following the redesign when the sample adult is completing the health insurance assessment on their own behalf.

While T-MSIS data quality varies by state and year, Medicaid and CHIP total enrollment is generally well reported when compared to an external benchmark.[70] Therefore, the T-MSIS data can be viewed as a gold standard and a suitable benchmark for assessing survey reported Medicaid coverage. The results of this analysis demonstrate that the linked NHIS-CMS T-MSIS files can be used to evaluate the accuracy of Medicaid enrollment information collected in NHIS; thus serving as a resource for public health researchers and survey methodologists in their analytic and design decisions.

6 Tables

Table VIII. Comparison of National Health Interview Survey report to T-MSIS record of Medicaid/CHIP coverage during interview month, 2016–2018

	T-MSIS record of full-scope or comprehensive Medicaid/CHIP coverage		
NHIS report of Medicaid/CHIP coverage	Yes	No	Total
Yes	14,177	2,187	16,364
No	4,216	74,209	78,425
Total	18,393	76,396	94,789

Overall agreement: 93.2%
Kappa statistic: 0.77 (SE=.003)
Sensitivity: 77.1%
Specificity: 97.1%
Positive predictive value: 86.6%
Negative predictive value: 94.6%
NOTES: NHIS is National Health Interview Survey. T-MSIS is Transformed Medicaid Statistical Information System. SE is standard error. Does not include 17-year-olds. Counts are unweighted.
SOURCE: Linked National Health Interview Survey and CMS Medicaid data

Table IX. Comparison of National Health Interview Survey report to T-MSIS record of any Medicaid/CHIP coverage during interview month, 2016–2018, ages 0–16 years

	T-MSIS record of full-scope or comprehensive Medicaid/CHIP coverage		
NHIS report of Medicaid/CHIP coverage	Yes	No	Total
Yes	7,033	929	7,962
No	1,645	12,621	14,266
Total	8,678	13,550	22,228

Overall agreement: 88.4%
Kappa statistic: 0.75 (SE=.005)
Sensitivity: 81.0%
Specificity: 93.1%
Positive predictive value: 88.3%
Negative predictive value: 88.5%
NOTES: NHIS is National Health Interview Survey. T-MSIS is Transformed Medicaid Statistical Information System. SE is standard error. Counts are unweighted.
SOURCE: Linked National Health Interview Survey and CMS Medicaid data

Table X. Comparison of National Health Interview Survey report to T-MSIS record of any Medicaid/CHIP coverage during interview month, 2016–2018, ages 18–64

	T-MSIS record of full-scope or comprehensive Medicaid/CHIP coverage		
NHIS report of Medicaid/CHIP coverage	Yes	No	Total
Yes	6,004	949	6,953
No	2,046	44,162	46,208
Total	8,050	45,111	53,161

Overall agreement: 94.4%
Kappa statistic: 0.77 (SE=.004)
Sensitivity: 74.6%
Specificity: 97.9%
Positive predictive value: 86.4%
Negative predictive value: 95.6%
NOTES: NHIS is National Health Interview Survey. T-MSIS is Transformed Medicaid Statistical Information System. SE is standard error. Counts are unweighted.
SOURCE: Linked National Health Interview Survey and CMS Medicaid data

Table XI. Comparison of National Health Interview Survey report to T-MSIS record of any Medicaid coverage during interview month, 2016–2018, age 65 and older

	T-MSIS record of full-scope or comprehensive Medicaid coverage		
NHIS report of Medicaid coverage	Yes	No	Total
Yes	1,140	309	1,449
No	525	17,426	17,951
Total	1,665	17,735	19,400

Overall agreement: 95.7%
Kappa statistic: 0.71 (SE=.01)
Sensitivity: 68.5%
Specificity: 98.3%
Positive predictive value: 78.7%
Negative predictive value: 97.1%
NOTES: NHIS is National Health Interview Survey. T-MSIS is Transformed Medicaid Statistical Information System. SE is standard error. Counts are unweighted.
SOURCE: Linked National Health Interview Survey and CMS Medicaid data

References

[1] National Center for Health Statistics, Office of Analysis and Epidemiology. The Linkage of National Center for Health Statistics Survey Data to Medicaid Enrollment and Claims Data - Methodology and Analytic Considerations. February 2019. Hyattsville, Maryland. https://www.cdc.gov/nchs/data/datalinkage/nchs-medicaid-linkage-methodology-and-analytic-considerations-508.pdf (accessed September 19, 2022)

[2] https://www.kff.org/wp-content/uploads/2013/01/8193.pdf (accessed September 19, 2022).

[3] https://www.medicaid.gov/medicaid/program-information/medicaid-and-chip-enrollment-data/report-highlights/index.html (accessed September 19, 2022).

[4] https://www.cms.gov/newsroom/fact-sheets/medicaid-facts-and-figures (accessed September 19, 2022).

[5] https://www.cms.gov/newsroom/press-releases/cms-office-actuary-releases-2019-national-health-expenditures (accessed September 19, 2022).

[6] https://www.ncsl.org/research/health/long-term-services-and-supports-faqs.aspx (accessed September 19, 2022).

[7] https://www.medicaid.gov/medicaid/benefits/behavioral-health-services/index.html (accessed September 19, 2022).

[8] https://www.kff.org/medicaid/issue-brief/medicaid-financing-the-basics/ (accessed September 19, 2022).

[9] 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 September 19, 2022).

[10] https://www.cms.gov/files/document/nhe-projections-2019-2028-forecast-summary.pdf (accessed September 19, 2022).

[11] https://www.medicaid.gov/sites/default/files/2019-12/list-of-eligibility-groups.pdf (accessed September 19, 2022).

[12] http://childwelfaresparc.org/wp-content/uploads/2014/10/Medicaid-to-26-for-Former-Foster-Youth7.pdf (accessed September 19, 2022).

[13] https://www.medicaid.gov/medicaid/benefits/early-and-periodic-screening-diagnostic-and-treatment/index.html (accessed September 19, 2022).

[14] https://www.macpac.gov/characteristics-of-key-medicaid-managed-care-spas-and-waivers/ (accessed September 19, 2022).

[15] https://www.macpac.gov/medicaid-101/waivers/ (accessed September 19, 2022).

[16] https://www.kff.org/medicaid/issue-brief/medicaid-waiver-tracker-approved-and-pending-section-1115-waivers-by-state/ (accessed September 19, 2022).

[17] https://www.cms.gov/Outreach-and-Education/American-Indian-Alaska-Native/AIAN/LTSS-TA-Center/info/1915-c-waivers-by-state (accessed September 19, 2022).

[18] https://www.medicaid.gov/about-us/program-history/index.html (accessed September 19, 2022).

[19] https://www.medicaid.gov/chip/eligibility/index.html (accessed September 19, 2022).

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

[21] https://www.medicaid.gov/state-overviews/scorecard/annual-medicaid-chip-expenditures/index.html (accessed September 19, 2022).

[22] https://www.kff.org/medicaid/state-indicator/total-chip-spending/ (accessed September 19, 2022).

[23] https://www.ssa.gov/OP_Home/ssact/title21/2103.htm (accessed September 19, 2022).

[24] https://www.medicaid.gov/dq-atlas/downloads/supplemental/3011_Final_Action_Status.pdf (accessed September 19, 2022).

[25] Nolan L, Barrett A, Nguyen L, Dowell J, Proctor K, and Parker J. "TAF Technical Documentation: Annual Demographic and Eligibility File." Baltimore, MD: CMS, 2021. https://resdac.org/sites/datadocumentation.resdac.org/files/2022-04/TAF_TechGuide_DE_File.pdf (accessed September 19, 2022)

[26] Christensen A, Arguello A, Hula L, et al. "TAF Technical Documentation: Claims Files." Baltimore, MD: CMS, 2021. https://resdac.org/sites/datadocumentation.resdac.org/files/2022-06/TAF-TechGuide-Claims-Files.pdf (accessed September 19, 2022)

[27] https://www.macpac.gov/wp-content/uploads/2016/03/Medicaid-Inpatient-Hospital-Services-Fee-for-Service-Payment-Policy.pdf (accessed September 19, 2022).

[28] 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/HealthCareFinancingReview/Downloads/CMS1192036dl.pdf (accessed September 19, 2022).

[29] https://www2.ccwdata.org/documents/10280/19002246/ccw-taf-rif-user-guide.pdf (accessed September 19, 2022)

[30] https://www.drugs.com/ndc.html (accessed September 19, 2022).

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

[32] Sayer, B. and C.S. Cox. *How Many Digits in a Handshake? National Death Index Matching with Less Than Nine Digits of the Social Security Number* in Proceedings of the American Statistical Association Joint Statistical Meetings. 2003. http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000144.pdf (Accessed September 19, 2022)

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

[34] Fellegi, I. P., and Sunter, A B. (1969), "A Theory for Record Linkage," JASA 40 1183-1210. DOI: 10.1080/01621459.1969.10501049 (Accessed September 19, 2022)

[35] Golden, C., et al., *Linkage of NCHS Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare Medicaid Services*. Vital Health Stat 1, 2015(58): p. 1-53. https://www.cdc.gov/nchs/data/series/sr_01/sr01_058.pdf (Accessed September 19, 2022)

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

[37] https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAX-Validation-Reports-Items/CMS1238402 (accessed September 19, 2022).

[38] https://www.macpac.gov/topics/dually-eligible-beneficiaries/ (accessed September 19, 2022).

[39] https://www.markfarrah.com/mfa-briefs/managed-medicaid-enrollment-trends-and-market-insights/ (accessed September 19, 2022).

[40] Filtered Managed Care Enrollment Summary (medicaid.gov) (accessed September 19, 2022).

[41] https://www.medicaid.gov/medicaid/managed-care/enrollment-report/index.html (accessed September 19, 2022).

[42] https://www.medicaid.gov/dq-atlas/landing/topics/single/map?topic=g8m81&tafVersionId=24 (accessed September 19, 2022).

[43] https://www.medicaid.gov/dq-atlas/downloads/supplemental/7011_Shared_Medicaid_IDs_2016.pdf (accessed September 19, 2022).

[44] https://www.medicaid.gov/status-of-t-msis-priority-items-1-12-of-july-2020/index.html (accessed September 19, 2022).

[45]https://www.medicaid.gov/status-of-t-msis-priority-items-1-23-of-july-2020/index.html (accessed September 19, 2022).

[46] https://www.medicaid.gov/medicaid/data-and-systems/macbis/tmsis/tmsis-blog/entry/54044 (accessed September 19, 2022).

[47] https://www.shadac.org/news/raceethnicity-data-cms-medicaid-t-msis-analytic-files-updated-february-2021-%E2%80%93-features-2018 (accessed September 19, 2022).

[48] https://resdac.org/ (accessed September 19, 2022).

[49] 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 September 19, 2022).

[50] 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 September 19, 2022).

[51] 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 September 19, 2022). Author manuscript available at: https://stacks.cdc.gov/view/cdc/109539 (Accessed September 19, 2022)

[52] 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. https://doi.org/10.2307/2289924 (Accessed September 19, 2022)

[53] 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. http://www.asasrms.org/Proceedings/papers/1990_056.pdf (Accessed September 19, 2022)

[54] 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 September 19, 2022).

[55] https://www.cms.gov/newsroom/fact-sheets/fact-sheet-medicaid-and-chip-t-msis-analytic-files-data-release (accessed September 19, 2022).

[56] Research Project to Understand the Medicaid Undercount. Phase IV Research Results: Estimating the Medicaid Undercount in the National Health Interview Survey (NHIS) and Comparing False-Negative Medicaid Reporting in NHIS to the Current

Population Survey (CPS). 2009. Available from: https://www.shadac.org/sites/default/files/publications/SNACC_Phase_IV_Full_Report.pdf

[57] Call K, Davern M, Klerman J, and Lynch V. Comparing errors in Medicaid reporting across surveys: evidence to date. Health Serv Res. 2013;48(2pt1):652–664

[58] Davern M, Klerman JA, Baugh DK, Call KT, and Greenberg GD. An examination of the Medicaid undercount in the current population survey: preliminary results from record linking. Health Serv Res. 2009 Jun;44(3):965-87

[59] Noon JM, Fernandez LE, and Porter SR. Response error and the Medicaid undercount in the current population survey. Health Serv Res. 2019 Feb;54(1):34-43. doi: 10.1111/1475-6773.13058. Epub 2018 Oct 1. PMID: 30270431; PMCID: PMC6338296.

[60] Davern M, Call KT, Ziegenfuss J, Davidson G, Beebe TJ, and Blewett L. Validating Health Insurance Coverage Survey Estimates: A Comparison of Self-Reported Coverage and Administrative Data Records, Public Opinion Quarterly, Volume 72, Issue 2, Summer 2008, Pages 241–259, https://doi.org/10.1093/poq/nfn013

[61] Mirel LB, Simon AE, Golden C, Duran CR, and Schoendorf KC. Concordance between survey report of Medicaid enrollment and linked Medicaid administrative records in two national studies. Natl Health Stat Report, 2014(72): p. 1-9

[62] https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questionnaires/NHIS/2016/english/qfamily.pdf [Accessed February 13, 2023]

[63] https://www.cdc.gov/nchs/nhis/health_insurance/hi_eval.htm [Accessed December 27, 2022]

[64] Nolan L, Barrett A, John J, Proctor K, and Parker J. "Identifying Beneficiaries with Full-Scope, Comprehensive, and Limited Benefits in the TAF." TAF DQ Brief #4151. Baltimore, MD: CMS, 2019. https://resdac.org/sites/datadocumentation.resdac.org/files/2021-01/4151_Scope_of_Benefits.pdf [Accessed September 26, 2024]

[65] Kincheloe JE, Brown ER, Frates J, Call KT, Yen W, and McRae JA. 2006. "Can We Trust Population Surveys to Count Medicaid Enrollees and the Uninsured? Health Affairs 25 (4): 1163–7

[66]https://resdac.org/cms-data/variables/eligibility-group-code-january [Accessed September 26, 2024]

[67] McHugh ML. Interrater reliability: The kappa statistic. Biochem Med (Zagreb) 22(3):276–82. 2012.

[68] Gordis L. Epidemiology. Philadelphia, PA: W.B. Saunders. 1996.

[69] https://www.kff.org/medicaid/issue-brief/status-of-state-medicaid-expansion-decisions-interactive-map/ [Accessed December 28, 2022]

[70] https://www.medicaid.gov/dq-atlas/landing/topics/single/map?topic=g1m7&tafVersionId=7 [Accessed June 26, 2023]