The Linkage of the National Center for Health Statistics Survey Data to 1996 – 2022 U.S. Department of Housing and Urban Development Administrative Data: Linkage Methodology and Analytic Considerations

Data Release Date: April 23, 2026

Document Version Date: April 23, 2026

Division of Analysis and Epidemiology

National Center for Health Statistics

Centers for Disease Control and Prevention

datalinkage@cdc.gov

Suggested Citation: National Center for Health Statistics, Division of Analysis and Epidemiology. The Linkage of the National Center for Health Statistics Survey Data to 1996 – 2022 U.S. Department of Housing and Urban Development Administrative Data: Linkage Methodology and Analytic Considerations, April 2026. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/linked-data/hud/index.html

Table of Contents

List of Acronyms

CMS, Centers for Medicare & Medicaid Services

DOB, date of birth

EM, expectation-maximization

ERB, Ethics Review Board

HCV, Housing Choice Voucher program

HUD, Department of Housing and Urban Development

LMF, Linked Mortality File

MEC, Mobile Examination Center

MF, Multi-family housing programs

HIC, Medicare Health Insurance Claim

MTW, Moving to Work program

NCHS, National Center for Health Statistics

NDI, National Death Index

NHANES, National Health and Nutrition Examination Survey

NHIS, National Health Interview Survey

PBS8, Project-based Section 8

PIC, Public & Indian Housing Information Center

PH, Public Housing program

PHA, Public Housing Agency

PII, Personally Identifiable Information

PW, Pair weight

RDC, Research Data Center

SS, Social Security

SSI, Social Security Income

SSN, Social Security number

SSN9, 9-digit Social Security number

SSN4, Last four digits of Social Security number

TRACS, Tenant Rental Assistance Certification System

1 Introduction

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

In a collaboration with the U.S. Department of Housing and Urban Development (HUD), the NCHS Data Linkage Program has been able to expand the analytic utility of the data collected from the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES) by augmenting it with housing assistance program data collected by HUD. This report will describe the linkage of data from 1999-2021 NHIS and 1999-March 2020 NHANES to 1996-2022 HUD administrative data. This linkage, collectively referred to as the Linked NCHS-HUD Data, creates a new data resource that can support a wide range of public health and policy evaluation studies focused on the relationship between housing and health.

This report includes a brief overview of the linked data sources, a description of the methods used for linkage, and analytic guidance to assist researchers when using the files. Detailed information on the linkage methodology is provided in Appendix I: Detailed Description of Linkage Methodology. More information about HUD housing assistance programs can be found in the companion document to these guidelines, "A Primer on HUD Programs and Associated Administrative Data" (A Primer on HUD Programs), or on the HUD website.[1] Additional documentation about the variables in the linked data files are available from the NCHS data linkage website.[2] Detailed information about the previous linkages of NCHS survey data and HUD administrative data have been published elsewhere.[3] [4]Although previous linkages have been conducted, the new linked data supersede the previous releases and should be used for all new analyses.

2 Date Sources

2.1 National Health Interview Survey (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

[1] U.S. Department of Housing and Urban Development, Office of Policy Development and Research (PD&R). HUD User. http://www.huduser.gov.

[2] National Center for Health Statistics. NCHS Data Linked to HUD Housing Assistance Program Files. https://www.cdc.gov/nchs/linked-data/hud/index.html.

[3] Lloyd PC, Helms VE, Simon AE, et al. Linkage of 1999–2012 National Health Interview Survey and National Health and Nutrition Examination Survey data to U.S. Department of Housing and Urban Development administrative records. National Center for Health Statistics. Vital Health Stat 1(60). 2017.

[4] National Center for Health Statistics. Division of Analysis and Epidemiology. The Linkage of the National Center for Health Statistics (NCHS) Survey Data to U.S. Department of Housing and Urban Development (HUD) Administrative Data: Linkage Methodology and Analytic Considerations, February 2022. Hyattsville, Maryland. https://www.cdc.gov/nchs/data/datalinkage/NCHS-HUD-Linked-Data-Methodology-and-Analytic-Considerations.pdf

United States. It is a multistage sample survey with primary sampling units of counties or adjacent counties, secondary sampling units of clusters of houses, tertiary sampling units of households, and finally, persons within households. It has been conducted continuously since 1957, and the content of the survey is periodically updated. Prior to 2007, NHIS traditionally collected full 9-digit Social Security Numbers (SSN) from survey participants. However, 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 in section 3.1. NHIS implemented its most recent content and structure redesign in 2019. For detailed information on the NHIS's content and methods, refer to the NHIS website, https://www.cdc.gov/nchs/nhis/index.html.

2.2 National Health and Nutrition Examination Survey (NHANES)

NHANES is a nationally representative cross-sectional survey designed to monitor the health and nutritional status of the civilian noninstitutionalized U.S. population. The NHANES sample is selected through a complex, multistage probability design. The sample design includes oversampling to obtain reliable estimates of health and nutritional estimates for population subgroups. The survey consists of interviews conducted in participants' homes and standardized physical examinations conducted in mobile examination centers. For detailed information about the Continuous NHANES contents and methods, refer to the NHANES website, https://www.cdc.gov/nchs/nhanes/index.html.

2.3 U.S. Department of Housing and Urban Development (HUD) Programs and Data

2.3.1 HUD Public and Assisted Housing Programs

The U.S. Department of Housing and Urban Development (HUD) is the primary federal agency responsible for overseeing domestic housing programs and policies. While HUD is responsible for administrating various housing and community development programs, the linkage with the 1999- 2021 NHIS and the 1999- 2020 NHANES focuses on HUD's three largest housing assistance programs: Housing Choice Vouchers (HCV), Public Housing (PH), and Multifamily (MF) programs. Persons and households participating in these program types are referred to as "HUD-assisted" in this document.

People living in HUD-assisted households are represented in HUD administrative data because they receive a rental subsidy or pay a below-market rent. HUD uses data about household characteristics (for example, household size and citizenship status, income, and expenses) to determine the amount of the rental subsidy under federal law. Generally, rental subsidies seek to reduce gross housing costs for the tenant to approximately 30% of household income, although program rules may allow for variations in that ratio. A HUD subsidy pays the remaining amount up to a specified limit that varies by program.

The HCV program is the federal government's largest housing assistance program, allowing families with lower incomes, older adults (persons 62 or older), and persons with disabilities to choose and lease safe and affordable housing. In the HCV program, housing assistance is tenant-based, meaning participants find their own housing in the private market. Participants are free to choose any housing that meets program requirements. In the NCHS-HUD linked data, the HCV program also includes several smaller programs including: Homeownership Vouchers, Project-Based Vouchers, Section 8 Moderate Rehabilitation, and the Section 8 Rental Certificate programs. Overall, among NHIS and NHANES participants that linked to HUD administrative data, slightly more than half were participating in an HCV program.

The public housing (PH) program was established to provide safe rental housing for eligible low-income families, the elderly, and persons with disabilities. HUD provides capital subsidies and operating subsidies to local Public Housing Agencies (PHAs) that manage public housing for eligible low-income residents. Unlike the HCV program, PH is project-based meaning tenants do not choose their housing but are instead assigned housing in a specific unit, building, or development. Overall, among NHIS and NHANES participants that linked to HUD administrative data, slightly less than one-third were participating in a PH program.

HCV and PH HUD program participants may also be participants in the Moving to Work (MTW) demonstration program. MTW provides PHAs the opportunity to design and test innovative, locally designed strategies that use Federal dollars more efficiently, help residents find employment, and increase housing choices for low-income families. Tenants participating in programs at MTW PHAs may need to verify their income and family composition less frequently than tenants in non-MTW HUD programs. These differences in program re-certification requirements were incorporated in the development of the linked NCHS - HUD administrative data files. (See Section 2.3.2 for more information on HUD administrative data).

The assisted multi-family housing (MF) program category in the linked NCHS–HUD data encompasses a number of separate, distinct HUD programs, including: Project-Based Section 8 (PBS8) the largest MF program, Section 221(d)(3) Below Market Interest Rate, Section 236 Multifamily Housing, Rental Assistance, Section 202 Supportive Housing for the Elderly Program, Section 202/162—Project Assistance Contract, Section 811 Supportive Housing for Persons with Disabilities, and Rent Supplement. Because each of the remaining MF programs lacked sufficient sample size on an individual basis in the linked file, they were combined into a single MF program category. In all MF programs, subsidies are paid directly to private property owners who provide a certain percentage of their housing units at affordable rates for low-income persons who qualify. MF program assistance is tied to the property, unlike tenant-based rental assistance programs (e.g., HCVs), and tenants cannot take their rental housing assistance subsidy elsewhere. Overall, among NHIS and NHANES participants that linked to HUD administrative data, slightly less than 40% were participating in an MF program.

2.3.2 HUD Administrative Data

HUD administrative data systems contain program participation data for recipients of HCV, PH, and MF programs for all states, the District of Columbia, and some territories (e.g., Puerto Rico and the U.S. Virgin Islands). The data collected through the administration of HUD's housing assistance programs are stored in two information management systems, the Public & Indian Housing Information Center (PIC) and the Tenant Rental Assistance Certification System (TRACS).

PIC contains household-level and person-level administrative records pertaining to persons and households participating in HUD's HCV and PH program types. The underlying forms used to capture information for these programs are the HUD-50058 and the HUD-50058MTW. The PIC data extract created for the NCHS-HUD data linkage was based on HUD's PIC point-in-time quarterly files, which capture a household's most recent transaction with HUD during the prior 18 months (except for Moving to Work (MTW) demonstration program participants, where 36 months is used as the threshold). A transaction refers to any activity for which a HUD form was completed (e.g., new admission to a HUD program, annual recertification, end of participation, etc.). These files are released four times a year (March, June, September, and December).

TRACS is a system to collect and maintain certified tenant data from owners and management agents of MF housing programs. The underlying form used to capture information for MF programs is HUD – 50059. The TRACS data extract created for the NCHS-HUD data linkage was based on TRACS point-in-time quarterly extracts from the TRACS production system. These data capture transactions occurring within the 18 months immediately prior to the date of the extract. Transactions with the same SSN, effective date, and transaction code were considered duplicates and removed.

To determine program overlap, HUD transactions collected from PIC and TRACS were used to create participation episodes for the final linked NCHS-HUD administrative data files. For more detailed information on the specific HUD data available on the NCHS-HUD linked data files, see Section 4.2.1.

HUD administrative records for MF program transactions that occurred between 1996 and 2022 were included in the linked datasets, and PH and HCV transactions occurring between 1999 and 2022 were included in the linked datasets.

For more information on HUD programs, their administration, and the PIC and TRACS data systems, please refer to A Primer on HUD Programs and Associated Administrative Data .

3 Linkage Methodology

3.1 Linkage Eligibility Determination

The linkage of NCHS-HUD data was conducted through a designated agent agreement between NCHS and HUD. Approval for the linkage was provided by NCHS' Research Ethics Review Board (ERB)[5] 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 name and date of birth, are considered linkage eligible. Linkage eligibility refers to the potential ability to link data from an NCHS survey participant to administrative data. Due to variability of questions across NCHS surveys, changes to PII collection procedures by the surveys over time, and changes in who is asked specific questions, criteria for NCHS-HUD linkage eligibility vary by survey and year.

For NHIS prior to 2007 and NHANES prior to 2009, a refusal by the survey participant to provide a 9-digit SSN (SSN9) was considered an implicit refusal for data linkage. However, NCHS observed an increase in the refusal rate for providing SSN, particularly for NHIS, which reduced the number of survey participants eligible for linkage.[6] In an attempt to address declining linkage eligibility rates, NCHS introduced new procedures for obtaining consent for linkage from survey participants. Research was also conducted to assess the accuracy of matching data from NHIS to the National Death Index (NDI)

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

[6] 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.

using partial SSN and other PII.[7] The research assessed algorithms using the last four and last six digits of SSN. The results provided support for changes in how NHIS collected SSN for linkage.[8]

Beginning in 2007, NHIS started requesting only the last four digits (SSN4) of SSN numbers. In addition, a short introduction before asking for SSN4 was added and participants who declined to provide SSN4 were asked for their explicit permission to link to administrative records without SSN. Also, at this time, the NCHS ERB determined that for the 2007 NHIS and all subsequent years, only primary respondents (sample adult and sample child) would be eligible for linkage to administrative records.

For the NCHS-HUD linkage, 1999-2006 NHIS participants were considered eligible for linkage if they:

- Did not refuse to provide SSN9, and
- Provided sufficient data elements for linkage.

Participants in the 2007-2021 NHIS were considered eligible for linkage if they:

- Provided SSN4 or an affirmative response to the follow-up question to allow linkage without SSN4, and
- Provided sufficient data elements for linkage.

For continuous NHANES, the informed consent procedures changed as well. NHANES continued to collect full nine-digit SSN 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 were given the option to provide only the last four digits.

For the NCHS-HUD linkage, 1999-2008 NHANES participants were considered eligible for linkage if they:

- Did not refuse to provide SSN9, and
- Provided sufficient data elements for linkage.

Participants in the 2009-2020 NHANES were considered eligible for linkage if they:

- Provided an affirmative response to the linkage consent question, and
- Provided the required data elements for linkage.

For both NHIS and NHANES, linkage was attempted only for survey respondents who provided consent for linkage, as described above, and who had at least two of the following three identifiers present:

- valid SSN[9]

[7] Sayer, B. and Cox, C.S. 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.

[8] Dahlhamer, J.M. and Cox, C.S., 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.

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

- valid date of birth (month, day, year)[10]
- valid name (first, middle, and last)[11]

Note that linkage eligibility is distinct from program eligibility, which defines whether a person meets the eligibility criteria for a specific government-administered or funded program. More information about HUD eligibility criteria is available from the HUD website at https://www.hud.gov/helping-americans/public-housing.

3.1.1 Match Rate Tables

Match rate tables providing NCHS-HUD linked samples sizes (number who were eligible for linkage, the number who were linked to HUD administrative data) and the percentage of total sample and eligible for linkage who were linked to HUD administrative program data for the total number of 1999-2021 NHIS and 1999-2020 NHANES participants, are available at https://www.cdc.gov/nchs/linked-data/hud/index.html.

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 that occurred prior to the survey participant's 18th birthday.

For example, a 2005 NHIS participant who was 15 years old at the time of interview can only be linked to HUD data for 2007 and earlier years (during which time the child was less than 18 years of age). This participant could not be linked to administrative records with dates after their 18th birthday, in this case dates beginning with 2008 and through later years.

3.3 Overview of Linkage

This section outlines steps that were used to link the NCHS survey data to the HUD enrollment data. For more detailed information on linkage methodology (see Appendix I).

Linkage-eligible NCHS survey participant records were linked to the HUD enrollment database using the following identifiers: SSN (9 digits or 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.

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

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

The NCHS survey participant records and the HUD enrollment database were linked using both deterministic and probabilistic approaches and were conducted separately for males and females[12].

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

2. Probabilistic linkage identified likely matches, or links, between all records. All records were probabilistically linked[13] and scored as follows:

    a. Formed pairs via blocking
    b. Scored pairs
    c. Modeled probability – assigned estimated probability that pairs are matches

3. Pairs were selected which were believed to represent the same individual between data sources (i.e., they are a match)

    a. Deterministic matches (from step 1) were assigned a match probability of 1
    b. Record selected from the probabilistic match (step 2) were assigned the model match probability

A score threshold was established for determining which NCHS survey participant were considered linked to the HUD enrollment database. For each NCHS participant record that was deemed a match, HUD extracted information from the PICS and TRACS systems and sent them to NCHS through a secure data transfer system.

4 Analytic Considerations

This section summarizes some key analytic issues for users of the linked NCHS survey data and HUD administrative records. It is not an exhaustive list of the analytic issues that researchers may encounter while using the linked NCHS-HUD 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). Users of the linked NCHS-HUD data files are encouraged to read "A Primer on HUD Programs and Associated Administrative Data" for additional information on HUD program and corresponding administrative data, including important analytic considerations.[14]

[12] Because first names are commonly associated with a person's sex, conducting the linkage separately for males and females helps to ensure independence and more appropriate weighting of name comparisons. Additionally, multiple part first and last names are more likely to be associated with females, which are handled differently when creating the linkage submission file. See Appendix I, Section 1 for additional information on the alternate record generation process for multiple part names.
[13] The probabilistic linkage methodology used is based on Fellegi, I. P., and Sunter, A B. (1969), "A Theory for Record Linkage," JASA 40 1183-1210.
[14] https://www.cdc.gov/nchs/data/datalinkage/primer-on-hud-programs.pdf

4.1 General Analytic Considerations for Linked Data

4.1.1 Access to the Restricted-Use Linked NCHS-HUD Data Files

To ensure confidentiality, NCHS provides safeguards including the removal of all personal identifiers from analytic linked files. Additionally, the linked data files are only accessible through the NCHS RDC network for approved applications. Researchers who wish to access the linked NCHS-HUD administrative data files must complete an RDC application. The RDC staff will review all submitted applications to determine if the proposed project is feasible and to identify any potential disclosure risks. More information regarding the NCHS RDC Network and the RDC application process is available from: https://www.cdc.gov/rdc/.

4.1.2 Merging Linked NCHS-HUD Data with NCHS Survey Data

To perform person-level analysis, the restricted-use Linked NCHS-HUD Data Analytic Files can be used in conjunction with the NCHS collected survey data (described above in Section 2.1 and 2.2). A unique survey participant identification variable is available on each file that allows analysts to merge survey data for survey participants with their information from the NCHS-HUD Linked Data files. The unique survey identifiers are survey-specific and may be constructed differently across survey years. Please refer to Appendix II: Merging Linked NCHS-HUD Files with NCHS Survey Data for guidance on identifying and constructing (if necessary) the appropriate identification variable for merging survey data and the NCHS-HUD Linked Data files.

4.1.3 Variables to Request in RDC Proposals

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

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

- Sample weights and design variables—these variables are needed to account for the complex design of the NCHS surveys. The names of the weights and design variables differ depending on which NCHS survey is being used. These can be identified using the documentation for each NCHS survey. As discussed below, NCHS recommends adjusting the sample weights to account for linkage eligibility bias. Linkage-eligibility adjusted weights are provided. However, researchers who wish to apply a different adjustment method should include the appropriate original sample weight(s) from the public use survey files.

- Demographic information about survey participants from the NCHS survey.

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

- Geography—Geography information is available on the administrative data for linked participants. However, there may be differences in the information available from the survey

and administrative data. It is recommended that users who require information on geography, request this information from the NCHS survey.

- Linked mortality data for NCHS surveys—Each of the NCHS surveys that have been linked to the HUD data have also been or will soon be linked to death information obtained from the NDI. The NCHS mortality files (LMFs) provide date and cause of death for each survey participant who has died. Researchers interested in analyzing linked mortality data with linked HUD data must specifically request the desired mortality variables in their RDC proposal. More information about the NCHS-NDI linked mortality files can be found at https://www.cdc.gov/nchs/linked-data/mortality-files/index.html.

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

4.1.4 Eligibility-adjusted Participant Survey 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 analysis.

One approach is to analyze linked data files using eligibility-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. 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*[15] and in *Assessing Linkage Eligibility Bias in the National Health Interview Survey*[16]

[15] 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.
[16] Aram, Jonathan et al., Assessing Linkage Eligibility Bias in the National Health Interview Survey. Vital and Health Stat 2, 2021(186).

The choice of which adjusted sample weight to use depends on the analysis and, more specifically, on the variables used in the analyses and the survey years included. Below are important considerations for the two surveys. NCHS has included eligibility-adjusted weights in the NCHS-HUD Weights file (see section 4.2.1).

For NHIS: Since all persons in the household sampled in the 1999-2006 NHIS were potentially eligible for linkage, eligibility-adjusted analyses of 1999-2006 NHIS should incorporate the person weights (FA_WGT_ADJ), or the sample adult weights (SA_FA_WGT_ADJ) (if analytic variables are based on sample adult file), or the sample child weights (SC_FA_WGT_ADJ) (if analytic variables are based on sample child file). As only sample adults or sample children were potentially eligible for linkage in the 2007-2021 NHIS, eligibility-adjusted analyses of 2007-2021 NHIS sample adult and sample child participants should either incorporate the adjusted sample adult weights (SA_FA_WGT_ADJ) or adjusted sample child weights (SC_FA_WGT_ADJ) respectively. For NHIS 2020, researchers are advised to review the NHIS 2020 Documentation for discussion of changes to field procedures in response to the COVID-19 pandemic, the three analytic data files available for sample adults, and the appropriate weight to use. The linked HUD file contains two additional weights for 2020 NHIS sample adults only: (1) sample adult longitudinal weight for analyses of data from 2019 and 2020 for the same individuals (SA_L_WGT_ADJ); and (2) sample adult partial weight for combining data from multiple years that include 2019 and 2020 (SA_P_WGT_ADJ).

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

If researchers wish to further adjust sample weights this can be done using the HUD_MATCH_STATUS variable to determine linkage eligibility from the NCHS-HUD weights file (see section 4.2.1).

4.1.5 Linked NCHS-HUD Link Probability Variable

For the survey data linked to HUD data, the probabilistic cut-off values used to determine which record pairs were considered a link (an inferred match) were set at a point that minimized both the type I error (false positives, or linked records that are not true matches) and the type II error (false negatives, or true matches that are not linked). For each candidate pair, the probability of link validity (PROBVALID) was computed and compared against a probability cut-off value to determine which pairs were links. For additional discussion on how PROBVALID was estimated, see Appendix I Sections 4.1 and 4.2.

In the NCHS-HUD linkage, NCHS used a probability cut-off value of 0.91 to determine final match status. Candidate pairs with a PROBVALID that exceeded the probability cut-off (i.e., PROBVALID>0.91) were considered linked.

Researchers can request access to PROBVALID (in the HUD weights file) in their RDC proposal to adjust linkage certainty by increasing the link acceptance cut-off scores or to conduct sensitivity analyses. For some analyses, it may be desirable to minimize type I error, which would be the result of using a value of PROBVALID closer to 1. Similarly, researchers may only want to include deterministic links, and could

restrict the analysis to records with PROBVALID=1. Note, the probability cut-off value cannot be decreased from 0.91 as pairs estimated with lower match probability are not made available to researchers.

4.2 Analytic Considerations for Linked HUD Data Files

4.2.1 Description of NCHS-HUD Linked Data Files

The NCHS-HUD linked data are comprised of the Transaction, Episode, Temporal Alignment, and Weights files. These files will be referenced in the remainder of the document. Variables found in each file can be referenced in the codebooks. The term "transaction" refers to any occurrence for which a HUD form is completed (e.g., new admission to a HUD program, annual recertification, end of participation, etc.). The term "episode" refers to a single continuous period of enrollment in a HUD program based on dates of HUD transactions. The Episode files are constructed from all transactions provided by HUD. The begin date of a participant's first episode is the effective date on their first transaction record. Subsequent episodes for the participant are identified based on the interval between the effective dates on their transaction records.

Transaction File

The transaction file contains a record for each transaction of the linked 1999-2021 NHIS-HUD and 1999-2020 NHANES-HUD participants. As noted previously, transactions for NHIS and NHANES child participants were removed during post-processing if the transaction occurred after their 18th birthday. The transaction file contains selected member and household attributes that are contained in HUD administrative systems. Each transaction includes an indicator for which episode it corresponds to and a count indicating the order of the transaction within the episode. Researchers can indicate how they would like to receive the transaction variable per episode, for example if they just want to include the first transaction or last transaction of the episode this should be indicated in the proposal requesting access to the linked files.

Episode Files

There are seven-episode files that contain start and end dates for participation episodes in various HUD programs based on the transaction data and assumptions about reasonable intervals between transactions. Most HUD recipients are required to recertify each year, and consequently, a transaction is expected each year. However, some HUD programs (for instance, the Moving to Work (MTW) Demonstration Program) have longer intervals between recertification. The episode files are useful primarily for longitudinal analysis related to the duration and timing of housing assistance episodes, and conditions or outcomes that may have preceded or followed such episodes.

The seven-episode files are:

- Episode File – Overall Universe: 1999-2021 NHIS and 1999-2020 NHANES participants who were linked with any transaction record in the HUD administrative data.
- Episode File – PH Universe: 1999-2021 NHIS and 1999-2020 NHANES participants who were linked with at least one PH program transaction in the HUD administrative data.
- Episode File – HCV Universe: 1999-2021 NHIS and 1999-2020 NHANES participants who were linked with at least one HCV program transaction in the HUD administrative data.
- Episode File – MTW PH Universe: 1999-2021 NHIS and 1999-2020 NHANES participants who were linked with at least one MTW PH program transaction in the HUD administrative data.

- Episode File – MTW HCV Universe: 1999-2021 NHIS and 1999-2020 NHANES participants who were linked with at least one MTW HCV program transaction in the HUD administrative data.
- Episode File – PBS8 Universe: 1999-2021 NHIS and 1999-2020 NHANES participants who were linked with at least one MF Project-Based Section 8 transaction in the HUD administrative data.
- Episode File – Other MF Universe: 1999-2021 NHIS and 1999-2020 NHANES participants who were linked with at least one Other MF program transaction in the HUD administrative data.

Appendix III: SAS Program to Create Participation Episodes provides the SAS program code used to create participation episodes.

Temporal Alignment File

The universe for the Temporal Alignment file includes 1999-2021 NHIS and 1999-2020 NHANES participants who were linked with any transaction record in the HUD administrative data. The temporal alignment file contains variables related to timing of HUD participation relative to the timing of the NHIS or NHANES interview and/or NHANES Mobile Examination Center (MEC) exam, such as: 1) indicator variables for receiving HUD-assisted housing on the date of the NCHS interview (or MEC examination, where appropriate), 2) the type of HUD-assisted housing received, and 3) the number of days between the interview and/or examination dates (NHANES MEC participants only) and the previous and/or next transactions. In addition, there are variables that indicate if the survey participant ever participated in the different HUD-assisted housing programs during the entire timespan of the administrative data.

Weights File

The universe for the Weights file includes all 1999-2021 NHIS and 1999-2020 NHANES participants. As mentioned previously, not all of the 1999-2021 NHIS and 1999-2020 NHANES participants are eligible for linkage. The variable HUD_MATCH_STATUS on the weights file indicates whether or not the survey participant was linkage eligible and if they linked to any HUD administrative records. In addition, the file includes NHIS and NHANES sample weights that have been adjusted for linkage eligibility. The Weights file contains a record for each 1999-2021 NHIS participant and each 1999-2020 NHANES participant. All participants who were ineligible for linkage (i.e., HUD_MATCH_STATUS equal to 9) are given an adjusted weight value of zero. Percentages related to linkage eligibility can be found in Tables 1 and 2 in the Match Rate Tables for NCHS-HUD linked data files (https://www.cdc.gov/nchs/linked-data/hud/index.html).
For more information on how the linkage eligibility-adjusted weights were created see Section 4.1.4.

Detailed descriptions for the complete list of variables contained in each of the NCHS-HUD linked data files can be found in the data dictionaries available on the NCHS Data Linkage website: https://www.cdc.gov/nchs/linked-data/hud/index.html.

4.2.2 Identification of Ever and Concurrent HUD-Assisted Survey Participants

4.2.2.1 Temporal alignment of survey and administrative data

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

4.2.2.2 Ever Received HUD-assisted Housing

To identify NCHS participants who live in HUD-assisted housing at any time during the administrative period (i.e., MF program transactions occurring during 1996 –2022, and HCV and PH transactions occurring during 1999 –2022), use the variable EVER_HUD on the Temporal Alignment File. To identify participants who ever lived in HUD-assisted housing through HCV, MF, and PH programs, use the variables EVER_MF, EVER_PH, and EVER_HCV, respectively.

4.2.2.3 Concurrent and Temporal Receipt of HUD-assisted Housing

The variables in the Temporal Alignment file can be used to identify concurrent HUD participation (i.e., participants who live in HUD-assisted housing at the time of their NCHS interview or examination, if applicable). Also included on the Temporal Alignment file are variables to identify participants who lived in HUD-assisted housing within a specific number of days before or after the survey interview or examination (TIME_A_INT, TIME_A_EXM, etc.). Because of disclosure risks, these count variables cannot be directly accessed by the researcher, but upon request, RDC staff can use them to derive categorical variables for researchers to use in the RDC. For example, to identify participants who were in HUD within 364 days (one year) of their NCHS interview, researchers may request in their RDC proposal that an indicator variable be created that identifies participants who lived in HUD-assisted housing within 364 days of their survey interview. Due to disclosure risks, derived variables based on the number of days before or after the survey will not be provided to researchers requesting episode files.

4.2.3. Analyses of Children in the Linked NCHS-HUD Data Files

As mentioned previously, administrative data for child survey participants generated after their 18th birthday are not available. This limitation affects groups two and three listed below for 1999-2021 NHIS or 1999-2020 NHANES child participants who lived in HUD-assisted housing during the 1996-2022 timeframe:

1. Child survey participants who only lived in HUD-assisted housing as children. There is no impact on this subgroup of children; all transactions are available.

2. Child survey participants who lived in HUD-assisted housing as children and adults. All transactions that occurred prior to the child's 18th birthday are available, but all transactions occurring on or after the child's 18th birthday are not available for release.

3. Child survey participants who lived in HUD-assisted housing only as adults. No transactions would be available for these participants in the NCHS-HUD linked data.

Researchers should keep in mind that for some survey years, adult survey participants may have HUD program participation available for transactions that occurred in the years prior to the interview when the participant was under 18 years of age. Researchers interested in performing analyses of children should take this into consideration.

4.2.4. Analyses of Rental Assistance Programs

Since a small number of HCV housing assistance programs provide homeownership vouchers, these programs are not technically "rental" assistance programs. Researchers using the linked files to specifically examine "rental" assistance programs should exclude transactions from the HCV

homeownership program. If researchers wish to broadly examine HUD assistance programs for low-income households, all transactions can be included. Researchers interested in examining only rental assistance programs should indicate this in their RDC proposal. NCHS will remove HCV homeownership vouchers from the requested file. More information about HCV homeownership vouchers is provided in Section 4.2.8. Variable Considerations and Data Anomalies.

For more detailed information on the types of housing-assistance programs administered by HUD and how HUD administrative data are collected, please refer to A Primer on HUD Programs and Associated Administrative Data.

4.2.5. Unit of Analysis

When using the NCHS-HUD linked files, the unit of analysis should be the participant, not the household. Survey participants, not households, were linked to HUD administrative data. Household-level analyses should not be done for several reasons. First, some members of a HUD household who were NCHS survey participants may not have been eligible for linkage; and will not be on the linked file. Second, transactions that occurred on or after the 18th birthday of child survey participants are not included in the linked files. Third, the membership of the HUD household may differ from that of the corresponding NCHS survey household.

4.2.6. Analytic Considerations for Episode Files

If the number of days between two transactions was within the recertification period (12 months for non-MTW recipients, 36 months for MTW recipients), the recipient was assumed to have been receiving assistance during that episode. If the number of days between two transactions was outside the recertification period, the end date was the previous transaction date.

There are two important considerations when using the episode files. First, transaction type was not taken into account when the episodes were created. The reason for using the number of days between the transactions rather than the type of transaction was that end of participation forms are not always submitted and requiring that an end of participation transaction define the end of an episode would bias concurrent predictions. As a result, given the way the episodes were defined, it is possible for an “end of transaction” to also appear as the start date of an episode.

It should be noted that researchers can use transaction type and end of participation dates to define their own episodes. However, this is not advisable without program expertise because, as noted above, this requires some assumptions about timing and may lead to misclassification.

The second consideration to keep in mind is that the overall episode file does not always align with the program-specific episode files. This is because the episodes in the overall episode file are created using the same algorithm as each program-specific episode file, which is based on the dates of transactions. The start and end dates are created irrespective of program type, which means that any two effective dates for two different programs may be the start and end date of a single episode. For program-specific analyses, data linkage staff recommend that the program-specific episode files be used in preference to the overall episode file. Episode files cannot be provided in conjunction with some variables on the temporal alignment file due to disclosure risks. Requests for variables from both episode and temporal alignment files will be subject to review.

4.2.7. MF Housing Program Data: Limitations and Considerations

Although HUD analysts generally do not treat the various MF subprograms as one composite category, a composite MF category was created for the NCHS-HUD linked files in addition to maintaining the MF subprograms. HUD does not recommend that researchers analyze MF subprograms without specialized expertise in these subprograms. HUD provides the following recommendations for analyzing MF programs in the linked data:

- If the research purpose is only to identify low-income individuals receiving HUD rental assistance, then use the pooled variable for MF.
- If the research purpose is to make program-specific policy recommendations related to MF housing, then acquire a comprehensive understanding of the various MF subprogram types and functions. Account for differences among the subprograms in the analysis, especially when inferences are drawn. Depending on the research question, it may be advisable to include only PBS8 participants in the analysis.
- In the linked data, the PBS8 program is the largest MF subprogram and the one most similar to the HCV and PH programs. Depending on the research question, it may be inadvisable to combine this program with the Section 236 or Section 221(d)(3) subprograms; doing so could lead to irrelevant and/or inaccurate results.
- Section 202 and Section 811 MF subprograms serve special populations- elderly households and disabled households. The differences between these populations and those of other HUD programs must be accounted for in the analysis, especially when inferences are drawn.

4.2.8. Variable Considerations and Data Anomalies

The HUD program data is collected for administrative purposes. It has been processed to be analytically useful for research purposes. This section outlines some of the variable considerations and data anomalies to be considered when using the linked data files.

4.2.8.1 Transaction File - HUD Program Variables

PROGRAM: Although HUD eligibility rules do not allow for subsidy payments for multiple HUD programs at the same point in time, a small number of individuals may appear to be recipients of more than one HUD program at the same time in the administrative data. Cases of dual program participation are rare but nonetheless exist in the linked data and indicate administrative data errors. Analysts must consider this potential discrepancy when conducting analyses using the linked data.

PROGRAM_TYPE: Some PROGRAM_TYPE variable codes in the transaction file have been re-categorized under the same overall PROGRAM variable. Program type codes for ‘Indian Housing’, ‘Certificate’, ‘Mandatory Conversion’, and ‘Moderate Rehabilitation’ have been recoded to the ‘Housing Choice Vouchers’ (VO) PROGRAM_TYPE category. The program type code for ‘Section 811 Project Rental Assistance Demo’ has been recoded to the ‘Section 202 PRAC (Project Rental Assistance Demo)’ (H7) PROGRAM_TYPE category.

4.2.8.2 Transaction File - Transaction Variables

Transactions with rare transaction codes were excluded from the NCHS-HUD linked data transaction file. Episodes of participation defined in the Episode files do not take into account the transaction type. Researchers interested in creating their own episodes using the type of transaction should understand

the recertification process for each HUD program. Recertification rules vary based on program and PHA participation in the MTW demonstration.

4.2.8.3 Transaction File - Disability Indicator and Count Variables

The transaction file includes one disability indicator variable (DISABLED_HOUSEHOLD) and one disability count variable (CHILD_DISABLED_CNT). Information on disability for HUD recipients is collected on Forms 50058 and 50059. These two HUD forms capture different definitions of disability which are defined according to program. It is important to note that the disability indicators are not related to the impairment variables (IMPRD_HEARING, IMPRD_MOBILITY, and IMPRD_VISUALLY), which are also on the transaction file.

Some of the disability variables are derived from other disability variables. For example, several conditions must be met in order to identify a household as disabled. A household is considered to be a HUD-disabled household if the head of household, spouse, and co-head are all less than 62 years of age and at least one of them is disabled. This is indicated by a household disability indicator (DISABLED_HOUSEHOLD) in the linked data. The child disability count variable is also derived from this variable as follows: CHILD_DISABLED_CNT is the count of all disabled household members who are under 18 years of age (including foster children).

4.2.8.4 Transaction File - Income Variables

The Transaction file has summary income variables that provide information about the income amounts and sources for the household as a whole. Some income codes are used to establish exclusions or deductions. When potential tenants apply for housing assistance, they must report all sources of income, except income for individuals explicitly excluded (i.e., live-in aides, foster children, and foster adults). Exclusions vary by HUD program.

The variable, TOT_A_INCOME provides total household income. TOT_A_INCOME is calculated by PHAs. If a researcher is interested in household income details such as majority income source of income, he/she should use MAJ_INCOME. The MAJ_INCOME is a categorical variable which is created by summing all income sources to the total annual household income amount after exclusions (including Pension, Social Security (SS)/Supplemental Security Income (SSI), then categorizing by source of that total income. The majority income is categorized by which income source comprises more than 50% of the total annual household income. If there is no majority income source, it is categorized as 'No Majority Source'.

Note that monetary values in the NCHS-HUD linked data files are not adjusted for inflation. General guidance from HUD's Economic and Market Analysis Division[17] is to use the Consumer Price Index (CPI)[18] when adjusting incomes and rents for comparability across time and geography. Due to fluctuations in the relationship between rent and utilities to gross rent, it is recommended to use 80% of the change in Rent of Primary Residence and 20% of the change in Fuels and Utilities when adjusting gross rent for inflation.

---

[17] The Economic and Market Analysis Division oversees HUD's field office economists who give the Department the capacity to have current, accurate, unbiased data and intelligence on local economic and housing market housing conditions/trends throughout the nation. More information can be found on this webpage: https://www.huduser.gov/portal/about/pdrdvsn_desc.html#econ_market_analysis.

[18] U.S. Bureau of Labor Statistics. Consumer Price Index. https://www.bls.gov/cpi/.

4.2.8.5 Transaction File - Total Household Expenses and Assistance Payments

The TOTAL_HOUSEHOLD_EXPENSES variable in the transaction file gives the total amount paid monthly by a household for expenses (i.e., rent and utilities). This variable may be inaccurate for participants in MTW programs, but the extent of the inaccuracy is unknown, and these calculations are provided by HUD only as an estimate for the researcher. Additionally, this variable was derived from multiple variables that are not available on the linked data files. For MTW records with negative values, these have been replaced with zeros. Assistance amounts are missing for PH programs because the subsidy is delivered via the operating fund and the capital fund, not to individual households. Calculations for assistance payments and total household expenses can be found in A Primer on HUD Programs and Associated Administrative Data.

4.2.9. Geocoded Data

Geocoded data for the linked participant's residence at the time of their survey interview are available through the RDC. However, it is important to note that although this level of geography is available, NHIS and NHANES samples are only representative at the regional and national level. For this reason, PHAs and private housing providers are not identified in the linked data.

Some NCHS surveys include a measure of urban/rural geographic location. Please refer to the survey documentation for information about available data. If the survey does not include the urban-rural classification of interest, it can be merged onto the file using state and county identifiers. An urban-rural classification recommended for use with NCHS surveys is the NCHS Urban-Rural Classification Scheme for Counties.[19] When requesting that an urban-rural classification scheme be merged onto the NCHS-HUD linked file, include state and county in the list of restricted variables and request the NCHS Urban-Rural scheme as an additional NCHS data source. State and county identifiers will be removed after the urban-rural codes are merged onto the linked file.

5 Additional Related Data Sources

Each of the NCHS surveys that have been linked to the HUD data have also been linked to death information obtained from the NDI. The linked mortality files provide the opportunity to conduct a vast array of outcome studies designed to investigate the association of a wide variety of health factors with mortality. For more information about the NCHS linked mortality files, please see the data linkage website: https://www.cdc.gov/nchs/linked-data/mortality-files/index.html.

NCHS survey data have also been linked to Centers for Medicare & Medicaid Services (CMS) Medicare and Medicaid enrollment and claims data. Researchers interested in analyzing information on HUD housing-assistance and health care utilization for persons also enrolled in Medicare may request variables from the NCHS-CMS Medicare Linkages, please see the data linkage website for more information: https://www.cdc.gov/nchs/linked-data/medicare/index.html.

[19] U.S. Centers for Disease Control and Prevention. NCHS Urban-Rural Classification Scheme for Counties. https://www.cdc.gov/nchs/data-analysis-tools/urban-rural.html

Researchers interested in analyzing information on HUD housing assistance and health care utilization for persons also enrolled in Medicaid may request variables from the NCHS-CMS Medicaid Linkages, please see the data linkage website for more information: https://www.cdc.gov/nchs/linked-data/medicaid/index.html.

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

Appendix I: Detailed Description of Linkage Methodology

1 NCHS and HUD Linkage Submission Files

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

- SSN (validated)[20] [21]
- DOB (month, day, and year)
- Sex
- Zip code and state of residence
- First, middle, and last name

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

- Date values: when invalid or outside of expected range, 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—name parts that contain specific keywords such as baby, infant, girl or boy are set to null
  - Names listed as Jane/John Doe
  - Removal of titles such as Mister, Miss, etc.
  - Removal of suffixes such as Junior, II, etc.
  - Removal of special text such as first name listed as “Void”

To increase the likelihood of finding a link, multiple or alternate submission records could be generated for each linkage eligible NCHS survey participant and HUD submission files based on variations of the linkage variables. Similar to the cleaning process, a more elaborate routine was used to generate alternate records involving the name fields. Alternate records were generated according to the following rules.

- Sex was missing. Two alternate records (one with male sex and the other with female) were created
- SSN with less than nine digits. A single alternate record was created where leading zeros were added to SSN values of length 7 or 8 to make a 9-digit SSN. Note, no alternate record was created if an invalid SSN would be created by adding 0’s.

[20] Complete SSN is considered valid if: 9-digits in length, containing only numbers, does not begin with 000, 666, or any values after 899, all 9-digits cannot be the same (i.e., 111111111, etc.), middle two and last 4-digits cannot be 0’s (i.e. xxx-00-xxxx or xxx-xx-0000), and is not 012345678. For some surveys and survey years, only the last 4-digits (SSN4) were collected from survey participants.

[21] If SSN missing or invalid, then SSN was extracted from their Medicare Health Insurance Claim (HIC) numbers, if provided. SSN was extracted from the Medicare HIC number only if the survey participant was identified as the primary claimant for Medicare benefits.

- Improbable date of birth. Age at time of survey was computed by subtracting the survey date (date last known alive) and birth date. If a date part was missing, age was computed by subtracting the year of the survey and the year of birth. Records with age greater than 114 had a single alternate record created,
    - If month and day were suspected of being imputed (ex. Jan 1st or June 15th), entire DOB was changed to missing
    - Otherwise, only year was changed to missing

- State of residence outside of U.S. and not in rest of world (RW) list. Alternate record was created with state code changed to missing

- ZIP code represents a different state. Using the ZIPSTATE() SAS function, state was imputed using the non-missing ZIP code. If the imputed state was different from the recorded state of residence, an alternate record using imputed state was created

- Multiple name parts and common nicknames (see below)

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

Table 1. Example of Alternate Record Generation using Name Fields

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

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

Submission files, which combined the cleaned and validated PII fields, were created separately for NCHS survey records and for HUD enrollment records. During this process, multiple submission file records were created for each survey participant/HUD enrollee to show all combinations of the recorded values for these fields. That is, if a survey participant had two states–of-residence recorded and three dates-of-birth recorded and each of the remaining fields had only one variant, then a total of six submission records would have been created for the survey participant (see Table 2 for example). Submission

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

Table 2. Example of Alternate Records Caused by Different PII Values

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

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

2 Deterministic Linkage Using Unique Identifiers

The deterministic linkage, which was the next step in the linkage process, used only the NCHS and HUD submission records that included a valid format SSN. Linkage eligibility is defined earlier in this report (see Section 3.1 Linkage Eligibility Determination). The algorithm performed two passes on the data, the first pass joining records when all 9-digits of the SSN matched and then for records where the last four digits of the SSN matched. After records had been linked using SSN, the algorithm validated the deterministic links by comparing first name, middle initial, last name, month of birth, day of birth, year of birth, ZIP code of residence, and state of residence. If the ratio of agreeing identifiers to non-missing identifiers was greater than 50% (1st pass using SSN-9) or greater than 2/3 (2nd pass using last 4 of SSN), the linked pair was retained as a deterministic match. In addition to the 2/3's agreement ratio, linked pairs in the 2nd pass were required to have at least 5 non-missing PII variables in agreement to be deemed a deterministic match. Of note, NCHS survey participants were excluded from the second pass (i.e., using the last 4-digits of SSN) if they were deterministically linked in the first pass. The collection of records resulting from the deterministic match is referred to as the 'truth source.'

3 Probabilistic Linkage

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

3.1 Blocking

Blocking is a key step in the probabilistic record linkage process. It identifies a smaller set of potential candidate pairs, eliminating the need to compare every single pair in the full comparison space (i.e., the Cartesian product). According to Christen, blocking or indexing, "splits each database into smaller blocks

according to some blocking criteria (generally known as a blocking key)."[22] Intuitively developed rules can be used to define the blocking criteria; however, for this linkage, variable values in the data being linked were used to inform the development of a set of blocking passes that efficiently join the datasets together (i.e., multiple, overlapping blocking passes are run, each using a different blocking key). By using these data to create an efficient block scheme (or set of blocking passes), a high percentage of true positive links were retained while the number of false positive links was significantly reduced. A supervised machine learning algorithm used the 'truth source' as the validation dataset and a sample of the NCHS survey and HUD 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."[23,24]

The machine learning algorithm learned 14 blocking passes to be used in the blocking scheme. Table 3 provides the PII variables that were assigned to each of the blocking passes and the PII variables that were used to score the potential links in each of the blocking passes. Note, the variables listed in the scoring key are all PII variables not used as a blocking variable. Further, if only the ZIP code of residence was used as a blocking variable, then state of residence was excluded from the list of scoring variables as it is implied to agree on all records.

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

[23] Michelson, M. and Knoblock, C.A. "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 .

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

Table 3. 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	First name, middle initial, last name, ZIP code of residence
3	Last name, first name, state of residence	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	First name, middle initial, day of birth, ZIP code of residence
5	First name, month of birth, year of birth, state of residence	Middle initial, last name, day of birth, ZIP code of residence
6	Last name, month of birth, day of birth, state of residence	First name, middle initial, year of birth, ZIP code of residence
7	First name, month of birth, day of birth, state of residence	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	First name, middle initial, month of birth, day of birth
13	Last name, day of birth, year of birth, state of residence	First name, middle initial, month of birth, ZIP code of residence
14	Month of birth, year of birth, state of residence, ZIP code of residence	First name, middle initial, last name, day of birth

3.2 Score Pairs

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

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

The pair scores were calculated on the 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 is the probability that the identifiers on a pair of records agree, given that the records represent the same person (i.e., the records are a match). M-probabilities were estimated separately within each individual blocking pass and were calculated for each of the identifiers used for scoring (Table 3). Within the blocking pass, pairs with agreeing SSN were used to calculate the M-probabilities, as these are assumed to represent the same individual. SSN agreement was defined as having 8 or more digits being the same for pairs with a full 9-digit SSN or the last 4-digits being the same for pairs with only a 4-digit SSN (ex. XXXXX9999). Further, to account for the alternate submission records generated during the creation of the submission files, the "best" agreement was taken for each of the scoring variables among the blocked records for each survey participant ID and HUD ID (see Tables 4 and 5 for example of alternate record summarization). Table 4 is an example of how the agreement flags for each of the scoring variables in blocking pass 10 are created. A value of 1 means the information in the variable is exactly matching, while a 0 means they are not. A value of "." (missing) means the scoring variable is missing for one or both data sources. Table 5 then represents how the multiple submission records in Table 4 are summarized into one record for each survey participant ID and HUD ID. If any of the identifiers agree across multiple records, they are flagged as agree (i.e., set to 1). The summarized records in Table V are then used to estimate the M-probabilities for each of the specific scoring variables.

Table 4. Example of Agreement Flags Using Blocking Pass 10

Person Identifiers		PII Agreement flags[1]				
Survey Participant ID	HUD ID	Middle Initial	Month of birth	Year of birth	ZIP Code	State of residence
1	1	1	0	1	0	.
1	1	.	1	1	0	0
1	1	1	0	1	0	0
2	2	1	0	1	0	0
3	789	1	1	.	0	1
3	789	0	1	0	1	1
3	789	.	1	0	1	.
3	789	0	0	1	1	1
3	322	1	0	1	1	1

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

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

Table 5. Example Showing Summarization of Blocked Record Pairs for M-Probability Estimation, based on Table 5 example

Person Identifiers		PII Agreement flags[1]				
Survey Participant ID	HUD ID	Middle Initial	Month of birth	Year of birth	ZIP Code	State of residence
1	1	1	1	1	0	0
2	2	1	0	1	0	0
3	789	1	1	1	1	1
3	322	1	0	1	1	1

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

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

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

- First/last initial agreement – used in the scoring process when only an initial was present in one or more of values (i.e., one from each of the two records being compared for a specific name variable)
- Jaro-Winkler Similarity Levels – this process is explained in greater detail in Section 3.2.2
- ZIP Code of residence – because ZIP codes are dependent on the state in which they are located, only the records where state of residence agreed were used in the computation of the ZIP code M-probability (i.e., if state was not in agreement, then it would be assumed that ZIP code would also not agree)

The U-probability is the probability that the two values for an identifier from paired records agree given that they were NOT a match. Similar to the M-probabilities, U-probabilities were calculated only for the

PII variables not included in the blocking keys and with the exception of first and last names, were computed within the blocking pass. The U-probabilities were computed using records where non-missing SSNs were not in agreement (defined as having less than 5 matching digits when records had a full 9-digit SSN and less than 4 matching digits for records with a 4-digit SSN). To avoid skewing U-probabilities in blocking passes that contained a high percentage of deterministic matches, assumed matches (i.e., records where SSN was not in agreement and had majority of the non-missing PII among scoring variables in agreement) were excluded prior to calculating the U-probabilities. For example, when computing the U-probability for day of birth in blocking pass 12, record pairs that did not agree on SSN that had a majority (i.e., greater than 50%) of the PII among first name, middle initial, and month of birth in agreement were excluded from the assumed non-matches. Even though SSN did not agree, these records were assumed to be probable links given that a majority of the PII between the NCHS survey and HUD submission records agreed.

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

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

For example, if for blocking pass 1, the state of residence code for FL appeared in 30,000 record pairings, agreed on 1,560 of those pairs, and the agreement rate for state of residence exceeded the 5th percentile, then the U-probability for Florida would have been computed as 1,560/30,000=0.052 or 5.2%. A 'catch-all' category was created for all identifier values that did not meet the above criteria. The U-probability of the 'catch-all' category was computed by dividing the total number of record pairs that agreed by the total number of record pairs being used to estimate the 'catch-all' category. Further, if there was no agreement in the 'catch-all' category, the U-probability would have been set to 0. To avoid a U-probability of 0, the 'catch-all' U-probability was computed by halving the minimum (i.e., lowest) U-probability among the individual value's U-probabilities. Further, if no individual value received a U-probability (i.e., all values assigned to 'catch-all') and there was no agreement, then the U-probability was set to 0.0001. For example, if the minimum U-probability among state of residence codes was 0.052 and there was no agreement among the catch-all records, the catch-all U-probability for state of residence would be 0.026 (0.052/2). If no state of residence code received a U-probability and there was no agreement, the U-probability for state of residence code would be 0.0001. The process for calculating U-probabilities for first and last name differs from these methods and is described in Section 3.2.2.

Lastly, an adjustment was made to the final U-probabilities to account for alternate records in the submission file. With the addition of each alternate record, the chance of agreement between the NCHS survey and HUD submission records increases. For example, a NCHS survey participant with different months of birth reported on two different participant records, has twice the chance of linking to a HUD submission record. Therefore, the U-probability for that participant's month of birth should represent the combined chance of agreement across both month values. Section 3.2.3 provides a detailed description of the methods used to adjust the U-probabilities to account for the additional alternate submission records.

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

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

Because of the large number of unique name values, it was impractical to compute U-probabilities specific to each name for each blocking pass (i.e., there would not be enough records available for it to be done accurately). Instead, U-probabilities were estimated using pairs generated by the Cartesian product of all records in the NCHS survey submission file and a simple random sample of 10% of records with non-missing name information of the HUD submission file.

Complete name tallies (separately, for first and last names) were then produced for the NCHS survey submission file. For each level of name on the file, 100,000 names were randomly selected from the HUD submission file 10% sample to compare. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. For each Jaro-Winkler level, the number of names in agreement of the 100,000 randomly selected HUD file names were then tallied.[25,26,27]

3.2.3 Adjustment of U-Probabilities for Alternate Submission Records

As previously mentioned in section 3.2.1, an adjustment was made to the U-probabilities to account for alternate submission records. The addition of unique values for an identifier increases the likelihood of a spurious linkage between records from the files being linked. Thus, the U-probabilities were adjusted to account for the increased probability of variable agreement (i.e., if records for the same person had multiple values for a variable, the chance of agreement with any compared record from the other file increases). Therefore, survey participants received an adjusted U-probability if they had identifier values that were different across their set of submission records. The adjusted U-probabilities were then applied to each record in the set of submission records that paired with a HUD administrative record. Lastly, the U-probability that is used to compute the agreement and disagreement weights (see Section 3.2.4) is the maximum between the original and adjusted U-probability (i.e., $U_{Max}=Max(U_{Original}, U_{Adjust})$).

Excluding first and last name and ZIP code of residence, the adjustment process began by identifying the unique set of values, and their U-probabilities, for each of the identifiers appearing in the scoring key (Table 3), for each survey participant. Because each value is assumed to be independent of the others, the adjusted U-probabilities were computed using the additive rule for probability as the

[25] 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.
[26] 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.
[27] Resnick, D., Mirel, L.B., 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.

summation of the individual value U-probabilities for each participant. That is, if a participant had three different month of birth values, the adjusted U-probability for month of birth was simply the summation of the three individual U-probabilities. Table 6 provides an example of the process used to compute the adjusted and maximum U-probabilities for month of birth.

Table 6. Example Showing Computation of the Adjusted and Maximum U-probability for Month of Birth

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

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

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

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

The first three columns of Table 6 show the unique values of month of birth and their corresponding U-probabilities (see Section 3.2.1) for survey participants 1, 2, and 3. The column titled "Adjusted U-Probability" is computed by totaling the individual probabilities in the third column for each participant. Finally, the maximum U-probability (last column), which was used to compute the agreement and disagreement weights (see Section 3.2.4), is the maximum value between the original and adjusted U-probability values.

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

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

where n is the number of unique state codes, m is the number of unique ZIP codes, and U_j is the U-probability for the j^{th} ZIP code. Table 7 provides an example of the process used to compute the adjusted U-probability for ZIP code of residence.

Table 7. Example Showing Computation of the Adjusted and Maximum U-probability for ZIP Code of Residence

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

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

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

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

The first four columns of Table 7 provide the Participant ID, state of residence, ZIP code of residence codes, and the corresponding U-probability for each ZIP code of residence for three survey participants. The adjusted U-probability (i.e., 5th column) is computed first by summing each individual U-probability within each state code and then taking the average of the summed values. The maximum U-probability (i.e., last column) is the max U-probability value between the original and adjusted ZIP-code of residence U-probabilities. Notice, for participants 8 and 25, the maximum U-probability value that was used for ZIP code 32011 and 39845, respectively, was the original U-probability. This was because the average U-probability across all state codes (column 5) did not exceed the original U-probability (column 4).

For first and last names, only the 85% Jaro-Winkler level U-probability was adjusted. The higher levels (i.e., 90, 95, and 100) were not adjusted because of the hierarchical method being used to compute each of the U-probabilities at those levels (i.e., 90 is dependent on 85, 95 is dependent on 90, and 100 is dependent on 95). Before the 85% level was adjusted, names that were similar to one another were combined into a single name field. This step is necessary to avoid 'double counting' names that are highly likely to match to the same name on the HUD administrative data file. Similarity in names was defined as having a Jaro-Winkler score between 0.95 and 1 (not inclusive at the upper bound) or if one name is fully contained within another (ex. Elizabeth and Eliza). If for example, a participant had two different names, Elizabeth and Elizabith (JW_{Score}=0.967), only one would be used to adjust the 85% Jaro-Winkler U-probability. The name that is selected was determined by whichever had the highest 100% Jaro-Winkler U-probability. Using the list of 'unduplicated' names, the adjusted U-probability for the 85% Jaro-Winkler level was computed as the summation of each of the individual U-probabilities for the participant. Table 8 provides an example of the methods used to compute the adjusted U-probabilities for the 85% Jaro-Winkler level, using first name as an example.

Table 8. Example Showing Computation of the Adjusted and Maximum U-probability for First Name

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

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

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

3.2.4 Calculate Agreement and Non-Agreement Weights

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

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

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

Agreement weights were only assigned to identifiers that had agreeing values. Similarly, non-agreement weights were only assigned to identifiers that had non-agreeing values. A non-agreement weight was always a negative value and reduced the pair weight score. It is important to note that if the M-probability was smaller than the U-probability (i.e., M<U), the pair score (see Section 3.2.5) was not adjusted according to the agreement/non-agreement weight. Because of the logarithmic function, having a M-probability that is smaller than the U-probability would have an inverse effect on the identifier agreement weights. That is, an agreement weight computed using a M-probability that was smaller than the U-probability would produce a negative weight, while the non-agreement weight would be positive. For example, if the M-probability for month of birth was 0.989 and the U-probability was 0.9999 then the agreement and non-agreement weights would be as follows,

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

Non-Agreement Weight (Identifier)

$$= \log_2\left(\frac{(1-M)}{(1-U)}\right) = \log_2\left(\frac{0.011}{0.0001}\right) = 6.781$$

3.2.5 Calculate Pair Weight Scores

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

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

First name and last name weights were assigned using Jaro-Winkler similarity scores described in Section 3.2.2. These scores ranged from 0 to 1, with 0 representing no similarity and 1 representing exact agreement. The weighting algorithm assigned all similarity scores 0.85 and below 0.85 a disagreement weight. The algorithm assigned all 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). 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 link probability, $P_{EM}(Match)$, for the potential matches in each blocking pass. The match probability represented the probability that a given link is a match. These probabilities in turn allowed the linkage algorithm to:

- Combine pairs across blocking passes (Pair-weights are specific to each blocking pass and are not comparable)
- Select a "best" record among survey participant's IDs that have linked to multiple administrative records
- Select final matches based on a probability threshold (discussed in the following section 4)

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

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

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

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

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

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

Continuing with the example from Step 1…

if for Pair 1 of blocking pass B, the pair-weight is 8.4, then $Odds_{1,B} = 2^{(8.4 + -5.61)} \approx 6.9$
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 $\widehat{N_{matches,B}}$ to be estimated. These estimated probabilities were then used to select the final matches, as described below in Section 4.

3.4 Adjustment for SSN Agreement

Up to this point, every pair generated through the probabilistic routine was assigned a value that estimates its probability of being a match. However, this estimate did not take SSN agreement into account. This was conducted as a separate step because for the other comparison variables, M- and U-probabilities were estimated based on probable matches or non-matches that were determined based on SSN agreement, and clearly this was infeasible for SSN itself.[28]

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 and HUD administrative record, the estimated probability was adjusted based on the last four digits of the SSN.[29]

When the last four digits of SSN[30] agreed (i.e., are exactly the same):

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

When the last four digits of SSN did not agree:

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

No adjustment was made for pairs that did not have an SSN on either the NCHS survey or HUD administrative record. So, for these pairs:

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

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

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

[30] 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 June 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

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

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, the percentage of them who were not true matches
- Type II Error: Among true matches, the percentage who were not linked

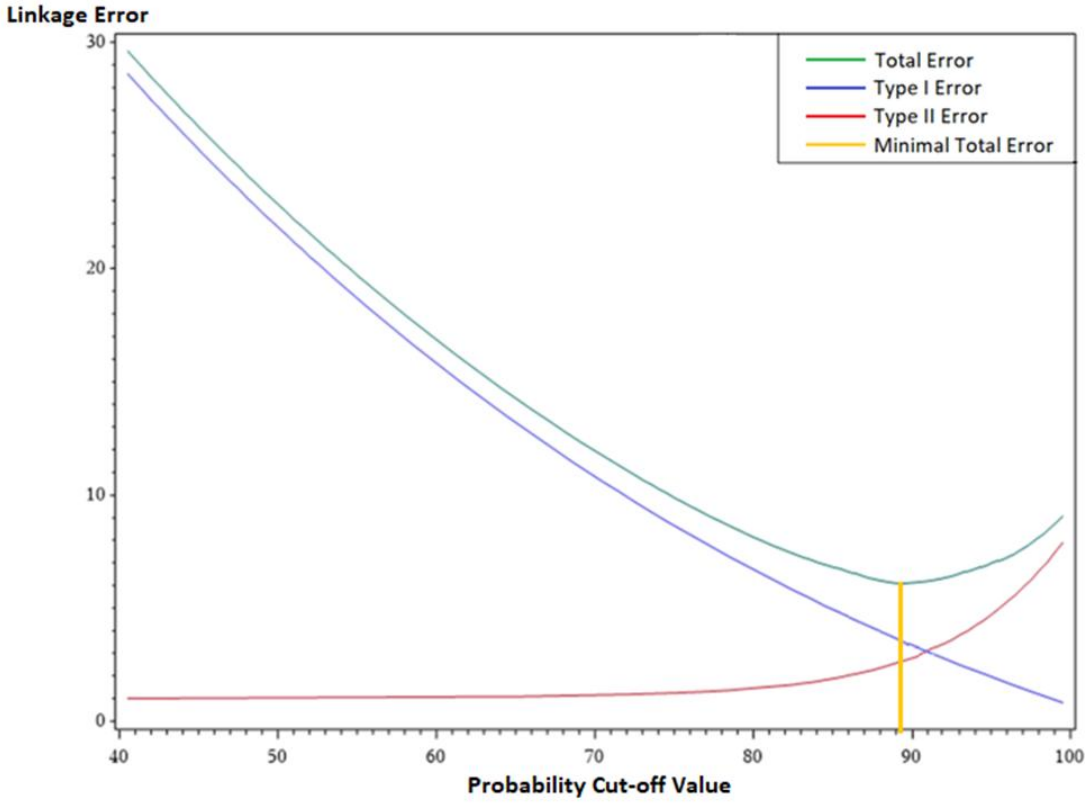
Because all records were included in the probabilistic linkage (i.e., even deterministic links), SSN agreement status (defined as seven or more matching digits for nine-digit SSN's and for SSN's that had only the last four digits, all four digits must match) was used to measure Type I error. Type I error for probabilistic links was measured as the total number of probabilistic links with non-agreeing SSN divided by the total number of probabilistic links with SSN available on both the NCHS survey and HUD 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. For example, if 40% of links were derived from the probabilistic method, this would reduce the estimated Type I error by the proportion of probabilistically determined linkages among all linkages. To further illustrate, if the Type I error rate for probabilistic links was estimated as 1.2%, then the estimated Type I error rate for the combined linkage process would be (0.40*0.012) = 0.0048 or 0.48%.

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

4.2 Set Probability Cut-off Value

One goal of record linkage is to have the lowest errors possible. However, as more pairs are accepted, pairs that were less certain to be matches but accepted as links increase the Type I error and decrease Type II error. And as less pairs are accepted, pairs that are more certain to be matches but not accepted as links decrease the Type I error and increase Type II error. The optimal trade-off between Type I error and Type II error is not known, but it can be assumed to be optimal when the sum of Type I and Type II error is at a minimum. For this reason, Type I and Type II error are estimated at various probability cut-off values and the one that showed the lowest estimate of total error is selected (see Figure 1). For this linkage, the probability cutoff was set to 0.91.

Figure 1: Illustrating linkage error by probability cut-off value
(Illustrative schematic not based on actual values)



4.3 Select Links Using Probability Threshold

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

4.4 Resolving NCHS participant IDs that Linked to Multiple HUD Enrollment Records

Due to the nature of administrative program data, it is possible that PII information may vary, due to PII changes over time or recording errors, among HUD enrollment records that represent the same person. In the 1999-2021 NHIS and 1999-2020 NHANES linked HUD files, a combined 32.2% of survey

participants were linked to more than one HUD enrollment record with the same HUD ID. In situations where a survey participant linked to more than one HUD enrollment record with different HUD IDs, and the PROVALID score calculated for each unique linked enrollment record exceeded the 0.91 cutoff value, all HUD ID matches were assumed to represent the same individual. In the 1999-2021 NHIS and 1999-2020 NHANES linked HUD files, about 4.2% of survey participants were linked to more than one HUD ID.

4.5 Computed Error Rates of Selected Links

Overall, the Type I and Type II linkage error rates for the NCHS survey-HUD Data linkage were 0.08% and 2.11%, respectively.

Appendix II: Merging Linked NCHS-HUD Files with NCHS Survey Data

The data provided on the 1999-2021 NHIS, 1999-2018 and 2017-March 2020 Pre-Pandemic NHANES linked HUD files can be merged with the NCHS restricted and public use survey data files using the unique survey specific public identification number (PUBLICID/SEQN).

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

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

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

NHIS, 1999-2003

Public-use

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

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

*The person number was called PX in the 1999-2003 NHIS and FPX in the 2004 (and later) NHIS; users may find it necessary to create an FPX variable in the 2003 and earlier datasets (or PX in later datasets).

SAS example: (note that the variables must be in character format for the concatenation)

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

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

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

R example:

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

NHIS, 2004

	Public-use		
Variable	Location	Length	Description
SRVY_YR	3-6	4	Year of interview
HHX	7-12	6	Household serial 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 a string variable)

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

R example:

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

NHIS, 2005 – 2018

	Public-use		
Variable	Location	Length	Description
SRVY_YR	3-6	4	Year of interview
HHX	7-12	6	Household serial 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 a string variable)

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

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

NHIS, 2019 – 2021

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

Note: The NHIS public-use files since the 2019 redesign do not contain a Person Number variable. To merge multiple NHIS public-use files, follow instructions provided in NHIS documentation. To merge to the linked data, concatenate the above variables from the Sample Adult file (RECTYPE=10 for all Sample Adults) or the Sample Child file (RECTYPE=20 for all Sample Children) to get the unique person identifier.

SAS example:
length PUBLICID $14;
PUBLICID = trim(left(SRVY_YR||HHX||RECTYPE));

Stata example: (note this will convert the variables to string variables)
egen PUBLICID = concat(SRVY_YR HHX RECTYPE)

R example:
Note that all PUBLICID components are read in as integers
df$PUBLICID<-paste0(sprintf("%04d", df$SRVY_YR), sprintf("%07d", df$HHX),sprintf("%02d", df$RECTYPE))

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

Item	Length	Description
SEQN	6	Participant identification number

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

Appendix III: SAS Program to Create Participation Episodes

Construction of the Episode Files

While a transaction is any occurrence for which a HUD form is completed (e.g., new admission to a HUD program, annual recertification, end of participation, etc.), an episode is a single continuous period of enrollment in a HUD program based on dates of HUD transactions. The Episode files are constructed from the Transaction file. The begin date of a participant's first episode is the effective date on their first transaction record. Subsequent episodes for the participant are identified based on the interval between the effective dates on their transaction records. The SAS program (described below) then cycles through each transaction's effective date and executes one of two actions:

1) treat the current transaction as part of the current episode and proceed to the next transaction, or
2) treat the current transaction as the start of a new episode, which then forces the previous transaction to be the end date of the previous episode.

The action implemented in the SAS code is determined by the number of days between each transaction as well as the HUD program type. The expected interval between any two transactions for a non-MTW recipient is one year. However, because PHAs are given 60 days leeway to submit reports, 425 days (one year plus 60 days) is used as the standard for determining if there has been a break in assistance. For most MTW PHAs, the expected interval between any two transactions for an MTW recipient is two years. However, MTW PHAs are given the flexibility to conduct recertification as infrequently as every three years. In the NCHS-HUD Linked Data, the estimated interval between any two transactions for the majority of MTW recipients is two years, again with a 60-day leeway; therefore, 790 days (two years and 60 days) is used as the standard.

If the interval between this date and the subsequent transaction's effective date is less than 425 days for non-MTW programs, or 790 days for MTW programs, it is assumed that the two dates are part of the same "episode" of participation. This continues until the interval between two effective dates is greater than 425 days for non-MTW programs, or 790 days for MTW programs. If the interval is greater than these durations, it is assumed that the two dates are from two distinct episodes of enrollment. In this situation, the effective date of the transaction immediately preceding the current transaction becomes the last date of the previous episode and the effective date of the current transaction becomes first date of the subsequent transaction.

The following SAS program was used to generate program participation episodes:

```
**************************************************************************
**************************************************************************** *** ***
***PURPOSE: CREATE EPISODE PERIODS FROM HUD TRANSACTION DATA *** *** ***** ***
**************************************************************************

**************************************************************************;

*@ACTION: INPUT RESTATE TRANSACTION FILE FROM HUD;
DATA NCHS_DATA;
  SET HUD_TRANSACTIONS_INT;
*@ACTION:RENAME PROGRAMS;
  IF PROGRAM EQ 'MTW HCV' THEN PROGRAM='MTW_HCV';
  IF PROGRAM EQ 'MTW PH' THEN PROGRAM='MTW_PH';
  IF PROGRAM EQ 'Other MF' THEN PROGRAM='OTHER_MF';
RUN;

%MACRO PERIODS (PGRM, PERIOD);
*@ACTION: SORT DATA BY IDS AND EFFECTIVE DATE;
PROC SORT DATA=NCHS_DATA OUT=ALL_DATA;
BY  NEW_HUD_ID PUBLICID EFFECTIVE_DATE;
*@ACTION: BREAK OUT BY NON MISSING PROGRAM TYPE;
%IF &PGRM GT  %THEN %DO;
  WHERE PROGRAM EQ "&PGRM";
%END;
RUN;
*@ACTION: CREATE INTERNAL EPISODE FILES BY PROGRAM TYPE;
DATA EPISODE_DATES_&PGRM._INT (KEEP=PUBLICID SEQN NEW_HUD_ID &PGRM._BEG_DATE1 -
&PGRM._BEG_DATE11 &PGRM._END_DATE1 - &PGRM._END_DATE11 SURVEY);
SET ALL_DATA;
BY NEW_HUD_ID PUBLICID;
*@ACTION:CREATE VARIABLES TO HOLD ALL OF THE PERIODS,PLUS BEGINNING AND ENDING DATES;
RETAIN HOLD_EFFDT PERIOD_1 - PERIOD_330 EPISODE_CNT TRANSACTION_CNT
       &PGRM._BEG_DATE1 - &PGRM._BEG_DATE11 &PGRM._END_DATE1 - &PGRM._END_DATE11;
EFFDT=EFFECTIVE_DATE;
ARRAY PERIODS (330) PERIOD_1 - PERIOD_330;
ARRAY BEGIN_DATE (11) &PGRM._BEG_DATE1 - &PGRM._BEG_DATE11;
ARRAY END_DATE (11) &PGRM._END_DATE1 - &PGRM._END_DATE11;
*@ACTION: LABEL VARIABLES;
LABEL
       PUBLICID            =      "NHIS PUBLIC USE ID"
       SEQN                =      "NHANES RESPONDENT SEQUENCE NUMBER"
       SURVEY              =      "SURVEY NAME"
       &PGRM._BEG_DATE1    =      "&PGRM. BEGIN DATE-EPISODE 1"
       &PGRM._BEG_DATE2    =      "&PGRM. BEGIN DATE-EPISODE 2"
       &PGRM._BEG_DATE3    =      "&PGRM. BEGIN DATE-EPISODE 3"
       &PGRM._BEG_DATE4    =      "&PGRM. BEGIN DATE-EPISODE 4"
       &PGRM._BEG_DATE5    =      "&PGRM. BEGIN DATE-EPISODE 5"
       &PGRM._BEG_DATE6    =      "&PGRM. BEGIN DATE-EPISODE 6"
       &PGRM._BEG_DATE7    =      "&PGRM. BEGIN DATE-EPISODE 7"
       &PGRM._BEG_DATE8    =      "&PGRM. BEGIN DATE-EPISODE 8"
       &PGRM._BEG_DATE9    =      "&PGRM. BEGIN DATE-EPISODE 9"
       &PGRM._BEG_DATE10   =      "&PGRM. BEGIN DATE-EPISODE 10"
       &PGRM._BEG_DATE11   =      "&PGRM. BEGIN DATE-EPISODE 11"
       &PGRM._END_DATE1    =      "&PGRM. END DATE-EPISODE 1"
       &PGRM._END_DATE2    =      "&PGRM. END DATE-EPISODE 2"
       &PGRM._END_DATE3    =      "&PGRM. END DATE-EPISODE 3"
       &PGRM._END_DATE4    =      "&PGRM. END DATE-EPISODE 4"
```

```
       &PGRM._END_DATE5    =      "&PGRM. END DATE-EPISODE 5"
       &PGRM._END_DATE6    =      "&PGRM. END DATE-EPISODE 6"
       &PGRM._END_DATE7    =      "&PGRM. END DATE-EPISODE 7"
       &PGRM._END_DATE8    =      "&PGRM. END DATE-EPISODE 8"
       &PGRM._END_DATE9    =      "&PGRM. END DATE-EPISODE 9"
       &PGRM._END_DATE10   =      "&PGRM. END DATE-EPISODE 10"
       &PGRM._END_DATE11   =      "&PGRM. END DATE-EPISODE 11"
;

*@ACTION: FORMAT THE DATE FIELDS;
FORMAT &PGRM._BEG_DATE1 - &PGRM._BEG_DATE11 &PGRM._END_DATE1 - &PGRM._END_DATE11
DATE.;
IF FIRST.PUBLICID THEN DO;

       HOLD_EFFDT=EFFDT;
*@ACTION: INITIALIZE FIELDS TO MISSING OR ZERO;
       DO J=1 TO 11;
              BEGIN_DATE (J)=.;
              END_DATE(J)=.;
       END;

       TRANSACTION_CNT=0;
       EPISODE_CNT=1;
       BEGIN_DATE(EPISODE_CNT)=EFFDT;

       DO I = 1 TO 330;
              PERIODS(I)=.;
       END;

END;
*@ACTION:  INCREMENT TRANSACTION COUNTER BY ONE;
TRANSACTION_CNT+1;
*@ACTION: CALCULATE PERIODS BETWEEN TRANSACTIONS;
PERIODS(TRANSACTION_CNT)=EFFDT-HOLD_EFFDT;
IF PERIODS(TRANSACTION_CNT) GT &PERIOD THEN DO;
       END_DATE(EPISODE_CNT)=HOLD_EFFDT;
       EPISODE_CNT+1;
       BEGIN_DATE(EPISODE_CNT)=EFFDT;
END;

HOLD_EFFDT=EFFDT;
*@ACTION: OUTPUT ONE RECORD PER ID;
IF LAST.PUBLICID THEN DO;
       END_DATE(EPISODE_CNT)=EFFDT;
OUTPUT;
END;
RUN;
*@ACTION: CREATE PUBLIC FROM INTERNAL VERSION;
DATA EPISODE_DATES_&PGRM._PUB(DROP=NEW_HUD_ID);
SET EPISODE_DATES_&PGRM._INT ;

RUN;
*@ACTION: SHOW CONTESNTS OF INTERNAL AND PUBLIC FILES;
PROC CONTENTS DATA=EPISODE_DATES_&PGRM._PUB varnum;
PROC CONTENTS DATA=EPISODE_DATES_&PGRM._INT varnum;
RUN;
%MEND PERIODS;
*@ACTION: RUN MACRO FOR ALL PROGRAM TYPE;
%PERIODS (HCV, 425); *NOTE:ONE YEAR PLUS TWO MONTHS;
%PERIODS (PH, 425);
%PERIODS (PBS8, 425);
%PERIODS (OTHER_MF, 425);
%PERIODS (MTW_HCV, 1155); *to use 38 MONTHS;
```

```
%PERIODS (MTW_PH, 1155);
%PERIODS (, 425);
```