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

Data Release Date: February 14, 2022

Document Version Date: April 28, 2023

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 (NCHS) Survey Data to U.S. Department of Housing and Urban Development (HUD) Administrative Data: Linkage Methodology and Analytic Considerations, February 2022. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data/datalinkage/NCHS-HUD-Linked-Data-Methodology-and-Analytic-Considerations.pdf

Table of Contents

List of Acronyms

CI, confidence interval

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

MBSF, Master Beneficiary Summary 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 collect, analyze, and disseminate timely, relevant, and accurate health data and statistics. NCHS products and services inform the public and guide program and policy decisions to improve our nation's health. In addition to collecting and disseminating the Nation's official vital statistics, NCHS conducts several population-based surveys and healthcare establishment surveys, including the National Health Interview Survey (NHIS), https://www.cdc.gov/nchs/nhis/index.htm (accessed October 12, 2021), and the National Health and Nutrition Examination Survey (NHANES), https://www.cdc.gov/nchs/nhanes/index.htm (accessed October 12, 2021). These surveys provide rich cross-sectional information on population characteristics and risk factors such as smoking, height and weight, health status, and socio-economic circumstances. Although the survey data collected provide information on a wide-range of health-related topics, they often lack information on longitudinal outcomes. Through its Data Linkage Program, NCHS has been able to enhance the survey data it collects by supplementing survey information with information from health-related administrative data sources. The linkage of survey and administrative data provides the unique opportunity to study changes in health status, health care utilization and expenditures, and other outcomes in specialized populations, such as people with low income and receiving housing assistance.

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 NHIS and NHANES by augmenting it with housing assistance program data collected by HUD. **This report will describe the linkage of data from the 1999 to 2018 NHIS and the 1999-2018 NHANES to 1996-2019 HUD administrative data.** This linkage, collectively referred to as the NCHS-HUD Linked Data, creates a new data resource that can support research studies focused on a wide range of health outcomes and the role of housing assistance programs as a social determinant of health and well-being.

This document describes the third and most recent linkage conducted between NCHS survey data and HUD administrative data. A brief overview of the data sources, a description of the methods used for linkage, analytic considerations, and description to linked data files, to assist researchers when using the data files are provided. Detailed information on the linkage methodology is provided in Appendix I: Detailed Description of Linkage Methodology. An evaluation of NCHS–HUD Linked Data program recipients compared with all HUD program recipients is included in Appendix II. 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," 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 linkage of NCHS survey data and HUD administrative data has

[1] U.S. Department of Housing and Urban Development, Office of Policy Development and Research (PD&R). HUD User. http://www.huduser.gov (accessed October 12, 2021).

[2] National Center for Health Statistics. NCHS Data Linked to HUD Housing Assistance Program Files. https://www.cdc.gov/nchs/data-linkage/hud.htm (accessed October 12, 2021).

been published elsewhere.[3] The new linked data supersede the previous release and should be used for all new analyses.

2 Background on Linked Files

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 United States. It is a multistage sample survey with primary sampling units of counties or adjacent counties, secondary sampling units of clusters of houses, tertiary sampling units of households, and finally, persons within households. It has been conducted continuously since 1957 and the content of the survey is periodically updated. Prior to 2007, NHIS traditionally collected full 9-digit Social Security Numbers (SSN) from survey participants. However, in attempt to address respondents' increasing refusal to provide SSN and consent for linkage, in 2007 NHIS began to collect only the last 4 digits of SSN and added an explicit question about linkage for those who refused to provide SSN. The implications of this procedural change on data linkage activities are discussed later in this report. For detailed information on the NHIS's content and methods, refer to the NHIS website, http://www.cdc.gov/nchs/nhis.htm (accessed October 12, 2021).

2.2 National Health and Nutrition Examination Survey (NHANES)

NHANES is a continuous, nationally representative survey consisting of about 5,000 persons from 15 different counties each year. For a variety of reasons, including disclosure concerns, the NHANES data are released on public-use data files in two-year increments. The survey includes a standardized physical examination, laboratory tests, and questionnaires that cover various health-related topics. NHANES includes an interview in the household followed by an examination in a mobile examination center (MEC). NHANES is a nationally representative, cross-sectional sample of the U.S. civilian, noninstitutionalized population that is selected using a complex, multistage probability design. For detailed information about the Continuous NHANES contents and methods, refer to the NHANES website, https://www.cdc.gov/nchs/nhanes/index.htm (accessed October 12, 2021).

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 to 2018 NHIS and the 1999 to 2018 NHANES focuses on HUD's three largest housing

[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.

assistance programs: Housing Choice Vouchers (HCV), Public Housing (PH), and Multifamily (MF) programs. Persons and households participating in these program types are "HUD-assisted."

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, 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 low-income families, elderly persons, and persons with disabilities to choose and lease safe and affordable housing. In the HCV program, housing assistance is tenant-based, meaning that participants find their own housing in the private market. Participants are free to choose any housing that meets program requirements and are not limited to units located in subsidized housing projects. In the NCHS-HUD linked data, the HCV program also includes the Homeownership Voucher, Project-Based Voucher, Section 8 Moderate Rehabilitation, and 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 MF program category in the linked NCHS–HUD data encompasses a number of separate, distinct HUD programs, including Project-Based Section 8 (or PBS8) Voucher Assistance in Multifamily Housing (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 half were participating in a MF program.

The PH program was established to provide safe rental housing for eligible low-income families, elderly people, and people with disabilities. HUD provides capital subsidies and operating subsidies to local Public Housing Agencies (PHAs) that manage public housing for eligible low-income residents. HUD also provides technical assistance to help PHAs plan, develop, and manage PH developments. Overall, among NHIS and NHANES participants that linked to HUD administrative data, about one-third were participating in a PH program.

2.3.2 HUD Administrative Data

HUD administrative data systems contain housing, income, and 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 on persons and households participating in HUD's HCV and PH program types. 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 (with the exception of 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.

TRACS is a system to collect and maintain certified tenant data from owners and management agents of MF housing programs. 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 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 June 30, 1996 and December 31, 2019 were included in the linked datasets, and PH and HCV transactions occurring between December 1, 1999 and December 31, 2019 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 (accessed October 12, 2021).

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).[4] The data linkage work was performed at NCHS.

Only a subset of 1999-2018 NHIS and 1999-2018 NHANES participants were eligible for linkage with the HUD administrative data. 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

[4] 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.

NCHS survey participant to administrative data. Criteria for NCHS-HUD linkage eligibility vary by survey and year due to variability of questions across NCHS surveys, changes to PII collection procedures by the surveys and changes in which survey participants are asked specific questions over time.

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.[5] 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) using partial SSN and other PII.[6] The research assessed algorithms using the last four and last six digits of SSN. The results were favorable and provided sufficient data to support changes in how NHIS collected SSN for linkage.[7]

Beginning in 2007, NHIS started requesting only the last four (instead of the full nine) digits of SSN (SSN4) 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) were eligible for data linkage.

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 2007-2018 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 NHANES, the informed consent procedures changed as well. SSN9 was consistently collected across the survey cycles for 1999-2018. 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.

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

[5] 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.

[6] 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.

[7] 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.

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

Participants in the 2009-2018 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.

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 (https://www.huduser.gov/portal/home.html).

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 1999-2018 NHIS and 1999-2018 NHANES participants, are available at https://www.cdc.gov/nchs/data/datalinkage/NCHS-HUD-Match-Rate-Tables-final.pdf.

3.2 Child Survey Participants

NCHS survey participants under 18 years of age at the time of the survey are considered linkage eligible if the linkage eligibility criteria described above are met and consent is provided by their parent or guardian. However, the consent provided by the parent or guardian does not apply once the child survey participant becomes a legal adult and there is no opportunity for NCHS to obtain consent to link the child participant's survey data to administrative data based on their adult experiences. As a result, in accordance with NCHS ERB guidance, NCHS only includes administrative data that were generated for program participation, claims and other events that occurred prior to the survey participant's 18th birthday on the linked data files provided to researchers.

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 data to the HUD enrollment database. For more detailed information on linkage methodology (see Appendix I).

Records from linkage-eligible NCHS participants were linked to the HUD enrollment database using the following identifiers: SSN, first name, last name, middle initial, month of birth, day of birth, year of birth, 5-digit ZIP code of residence, state of residence, and sex.

The NCHS survey participant records and the HUD enrollment database were linked using both deterministic and probabilistic approaches. For the probabilistic approach, scoring was conducted according to the Fellegi-Sunter method.[8] Following this, a selection process was implemented with the goal of selecting pairs that represented the same individual between the data sources. The following three steps were applied to determine linked records:

1. Deterministic linkage joins records on exact SSN, with links validated by comparing other identifying fields (i.e., first name, last name, day of birth, etc.)
2. Probabilistic linkage identified likely matches, or links, between all records. All deterministically linked pairs (from Step 1) were assigned a probabilistic probability of 1; other records were linked and scored as follows:
    a. Formed pairs via blocking
    b. Scored pairs
    c. Modeled probability – assigned estimated probability that pairs are links
3. Pairs were selected that were believed to represent the same individual between data sources (i.e., they are a link)

For each NCHS participant record that was linked, 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.[9]

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 made available in secure facilities for approved research projects. Researchers who wish to access the linked NCHS-HUD administrative data files must submit a research proposal to the NCHS Research Data Center (RDC) to obtain permission to access the restricted use files. All researchers must submit a research proposal to determine if their projects are feasible and to gain access to these restricted data files. The proposal provides a framework which allows RDC staff to identify potential disclosure risks. More information regarding RDC and instructions for submitting an RDC proposal are available from: https://www.cdc.gov/rdc/ (accessed October 12, 2021).

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

[9] https://www.cdc.gov/nchs/data/datalinkage/primer-on-hud-programs.pdf (accessed October 12, 2021)

4.1.3 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 III: 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.

4.1.4 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 RDC for merging with the requested restricted variables from NCHS surveys and for use with the variables 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 also need to be specifically requested as part of a researcher's application to RDC. Staff in the RDC verify the full list of variables (restricted and public-use) and check for potential disclosure risk.

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.
- 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 linked NDI mortality files 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/data-linkage/mortality.htm (accessed October 12, 2021).
- 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.

4.1.5 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.[10]

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 (ADJ_PERWT), or the sample adult weights (ADJ_SAWT) (if analytic variables are based on sample adult file), or the sample child weights (ADJ_SCWT) (if analytic variables are based on sample child file). As only sample adults or sample children were potentially eligible for linkage in the 2007-2018 NHIS, eligibility-adjusted analyses of 2007-2018 NHIS sample adult and sample child participants should either incorporate the adjusted sample adult weights (ADJ_SAWT) or adjusted sample child weights (ADJ_SCWT) respectively.

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).

[10] 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.

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).

4.1.6 Linked NCHS-HUD Link Probability Variable

Data linkages engender some uncertainty over which pairs represent true matches. 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. The PROBVALID cutoff is the threshold that produces the lowest total error (both type I and type II). However, because there are concerns that using pairs with low PROBVALID might be inappropriate for certain analyses of linked records, PROBVALID = 0.9225 was established as the lowest threshold that will be used for the acceptance of links into datasets made available for external researchers.

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.

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 1999-2018 NHIS-HUD and 1999-2018 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 select 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: 1999-2018 NHIS and 1999-2018 NHANES participants who were linked with any transaction record in the HUD administrative data.
- Episode File – PH Universe: 1999-2018 NHIS and 1999-2018 NHANES participants who were linked with at least one PH program transaction in the HUD administrative data.
- Episode File – HCV Universe: 1999-2018 NHIS and 1999-2018 NHANES participants who were linked with at least one HCV program transaction in the HUD administrative data.
- Episode File – MTW PH Universe: 1999-2018 NHIS and 1999-2018 NHANES participants who were linked with at least one MTW PH program transaction in the HUD administrative data.
- Episode File – MTW HCV Universe: 1999-2018 NHIS and 1999-2018 NHANES participants who were linked with at least one MTW HCV program transaction in the HUD administrative data.
- Episode File – PBS8 Universe: 1999-2018 NHIS and 1999-2018 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: 1999-2018 NHIS and 1999-2018 NHANES participants who were linked with at least one Other MF program transaction in the HUD administrative data.

Appendix IV: 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 1999-2018 NHIS and 1999-2018 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 1999-2018 NHIS and 1999-2018 NHANES participants. As mentioned previously, not all of the 1999-2018 NHIS and 1999-2018 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 1999-2018 NHIS participant and each 1999-2018 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 Table 1, and the Match Rate Tables for NCHS-HUD linked data file (NCHS-HUD match rate tables (cdc.gov)).
For more information on how the linkage eligibility-adjusted weights were created see Section 4.1.5.

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/data-linkage/hud-restricted.htm.

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 the dates June 30, 1996 – December 31, 2019, and HCV and PH transactions occurring during the dates December 1, 1999 – December 31, 2019), 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 1999-2018 NCHS-HUD Linked Data Files

As mentioned previously, administrative data for child survey participants generated after their 18th birthday are not available. This limitation impacts the latter two of the following three groups of 1999-2018 NHIS or 1999-2018 NHANES child participants who lived in HUD-assisted housing during the 1996-2019 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 (accessed Oct 12, 2021).

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. While one household member's living in HUD-assisted housing directly affects all members in a household, 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 PBS8participants in the analysis.

- In the linked data, the PBS8program 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: Individuals may be recipients of more than one HUD program at the same point in time. Cases of dual program participation are rare but nonetheless occur in the linked data and indicate errors in the administrative data. 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 ‘Section202 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. As described previously, the 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 variables (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 is to use the Consumer Price Index (CPI)[11] 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.

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 (accessed Oct 12, 2021).

4.2.9. Geocoded Data

Though the original file received from HUD contained a very detailed level of geographic

[11] U.S. Bureau of Labor Statistics. Consumer Price Index. https://www.bls.gov/cpi/. Accessed January 11, 2022.

information (i.e., address), the Transaction file available for analysis through the NCHS RDC does not contain this detailed level of geographic information.

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, whereas others do not. 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.[12] 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 or will soon be linked to death information obtained from the NDI. The linked NDI mortality files provides the opportunity to conduct a vast array of outcome studies designed to investigate the association of a wide variety of health factors with mortality. More information about the NCHS NDI linked mortality files can be found at https://www.cdc.gov/nchs/data-linkage/mortality.htm (accessed January 14, 2022).

NCHS has also previously linked to Center 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/data-linkage/medicare.htm (accessed October 12, 2021). 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/data-linkage/medicaid.htm (accessed October 12, 2021).

Data users may request variables from the Linked CMS Medicare, CMS Medicaid, or Linked NDI files (if mortality is the outcome of interest) 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 III).

[12] U.S. Centers for Disease Control and Prevention. NCHS Urban-Rural Classification Scheme for Counties. https://www.cdc.gov/nchs/data_access/urban_rural.htm. Accessed January 14, 2022.

Appendix I: Detailed Description of Linkage Methodology

1 NCHS and HUD Linkage Submission Files

A submission file is a dataset specially prepared for submission to the linkage analysis process, by having all necessary variables and records correctly formatted for this. Submission files, which contained the cleaned and validated PII fields, were separately created for NCHS survey records and for HUD enrollment records. To accomplish this, there were an initial series of processes that performed various data cleaning routines on the PII fields within each of the separate files containing NCHS survey and HUD administrative records, prior to their linkage. Of note, processing was conducted separately for NCHS and HUD records. The following PII fields were individually processed and output to its own file (i.e., there were separate files for SSN, DOB, name, etc., each record showing a possible value for that field for each survey participant or HUD enrollee):

- SSN (validated)[13] [14]
- DOB (month, day, and year)
- Sex
- 5-Digit ZIP code and state of residence
- First, middle, and last name

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

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

To increase the likelihood of finding a link, multiple or alternate submission records were used for each linkage eligible NCHS survey participant based on variations of the linkage variables. HUD records could be matched to any or all of the submission records created for a survey participant. Similar to the cleaning process, a more elaborate routine was used to generate

[13] 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.

[14] 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.

alternate records involving the name fields. For survey participants with multiple name parts, common nicknames, and for common Hispanic and Asian names, additional records were generated using each individual piece as a possible name value. For example, the name “Beth” may be a nickname for a formal name like “Elizabeth.” In this situation, a record for “Beth” and a record for “Elizabeth” were created and submitted for linkage. NCHS created a common nickname lookup file which was used to generate a second record replacing the nickname with the formal name. Table 1 below provides two examples of how multiple part name information was used to generate alternate records, using hypothetical data. For survey participant A, the first name was used to generate multiple records, and for survey participant B, the last name was used.

Table 1. Example of Alternate Record Generation using Name Fields

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

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

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

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[15]. Linkage eligibility is defined earlier in this report (see Section 3.1 Linkage Eligibility Determination). The algorithm performed two passes on the data, first checking for full 9-digit SSN agreement and then for records where the last 4-digits of the SSN agreed. After records had been 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. 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 data linkage expert Peter Christen, blocking or indexing, "splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key)."[16] Intuitively developed rules can be used to define the blocking criteria; however, for this linkage, variable values in the data being linked were used to inform the development of a set of blocking passes that efficiently join the datasets together (i.e., multiple, overlapping blocking passes are run, each using a different blocking key). By using these data to create an efficient block scheme (or set of blocking passes), a high percentage of true positive links were retained while the number of false positive links were significantly reduced. A supervised machine learning algorithm used the 'truth source' as the validation dataset and a sample of the NHCS and HUD submission records as training data. For

[15] 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.

[16] 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 (accessed October 12, 2021).

more detailed information on the supervised machine learning algorithm used please refer to "Learning Blocking Schemes for Record Linkage."[17,18]

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 be in agreement on all records.

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

[17] 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 (accessed October 12, 2021).

[18] Campbell, S.R., Resnick, D.M., Cox, C.S., & Mirel, L.B. (2021). Using supervised machine learning to identify efficient blocking schemes for record linkage. Statistical Journal of the IAOS, 37(2), 673–680. https://doi.org/10.3233/SJI-200779 (accessed October 12, 2021).

13	Last name, day of birth, year of birth, state of residence, sex	First name, middle initial, month of birth, ZIP code of residence
14	Month of birth, year of birth, state of residence, ZIP code of residence	First name, middle initial, last name, day of birth, sex

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
- Sex
- State of Residence
- ZIP Code (conditional on state agreement)

3.2.1 Calculate M- and U- Probabilities

The **M-probability** – the probability that the values of identifiers on a pair of records agree, given that the records represent the same person (i.e., the records are a match) – was estimated separately within each individual blocking pass. M-probabilities were calculated for each of the identifiers not used in the blocking key (Table 3). Within the blocking pass, pairs with agreeing SSN (defined as 8 or more digits being the same) were used to calculate the M-probabilities, as these are assumed to represent the same individual.

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

- First/last initial agreement – used in the scoring process when only an initial was present in the name field
- Jaro-Winkler Similarity Levels – this process is explained in greater detail in Section 3.2.2

- Last name is conditional on sex – because women frequently change their maiden name to their spouse's last name after marriage (or may change back to maiden in event of divorce/widowing), this resulted in a lower agreement last name M-probabilities for the female population and was taken into consideration when computing corresponding agreement and non-agreement weights.
- ZIP Code of residence – because ZIP codes are dependent on the state in which they are located, only pairs of 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 do NOT represent the same person (i.e., the records are not a match). With the exception of first and last names, these probabilities were calculated within each block, using records where non-missing SSNs were not in agreement (i.e., less than 5 digits are the same).

Similar to the M-probabilities, U-probabilities were only calculated for the non-blocking variables. However, for this linkage, the U-probabilities were calculated for each value (level) of a variable. However, for first and last name, the U-probabilities were not calculated exactly in the same manner, and the method used for them is described in Section 3.2.2.

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

Similar to the M-probability, Jaro-Winkler levels (0.85, 0.90, 0.95, and 1.00) were calculated for use in the U-probability computation. Because agreement levels fall over a range, first and last name U-probabilities were computed for each Jaro-Winkler score level. Since there are a plethora of possible values for first and last name (i.e., one for each possible name), it was impractical to compute U- probabilities for a specific name for each blocking pass (i.e., there would not be enough records available for it to be done accurately). Instead, U-probabilities were estimated using pairs generated by the Cartesian product of all records in the NCHS survey submission file and a simple random sample of 1% 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 1% sample to compare to it. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. The number of names in agreeance of the 100,000 randomly selected HUD file names that agreed at that level for each name were then tallied.[19,2021]

[19] Jaro M. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J Am Stat Assoc. 1987 Jan 01;406:414-420.

[20] Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association. 1990. 354-9.

[21] Resnick, D., Mirel, L.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 (accessed October 12, 2021).

3.2.3 Calculate Agreement and Non-Agreement Weights

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

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

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

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

3.2.4 Calculate Pair Weight Scores

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

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

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

3.3 Probability Modeling

A probability model, developed from a partial expectation-maximization (EM) analysis, was applied individually to each of the blocks in the blocking scheme. Each model estimated a 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)

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

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

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

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

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

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

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

   Continuing with the example from Step 1…

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

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

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

Continuing with the example…

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

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

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

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

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

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

3.4 Adjustment for SSN Agreement

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

To remedy this, before the algorithm adjudicated the matches against the probability threshold, one final adjustment was made to the match probabilities (for probabilistic pairs). For pairs that had an SSN on both the NHCS and HUD administrative record, the estimated probability was adjusted based on the last four digits of the SSN.[23]

When the last four digits of SSN[24] agreed (i.e., are exactly the same):

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

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

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

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

When the last four digits of SSN did not agree:

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

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

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

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

4.1 Estimating Linkage Error to Determine Probability Cutoff

Subsequent to performing the record linkage analysis an error analysis was performed. There are two type of errors that were estimated:

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

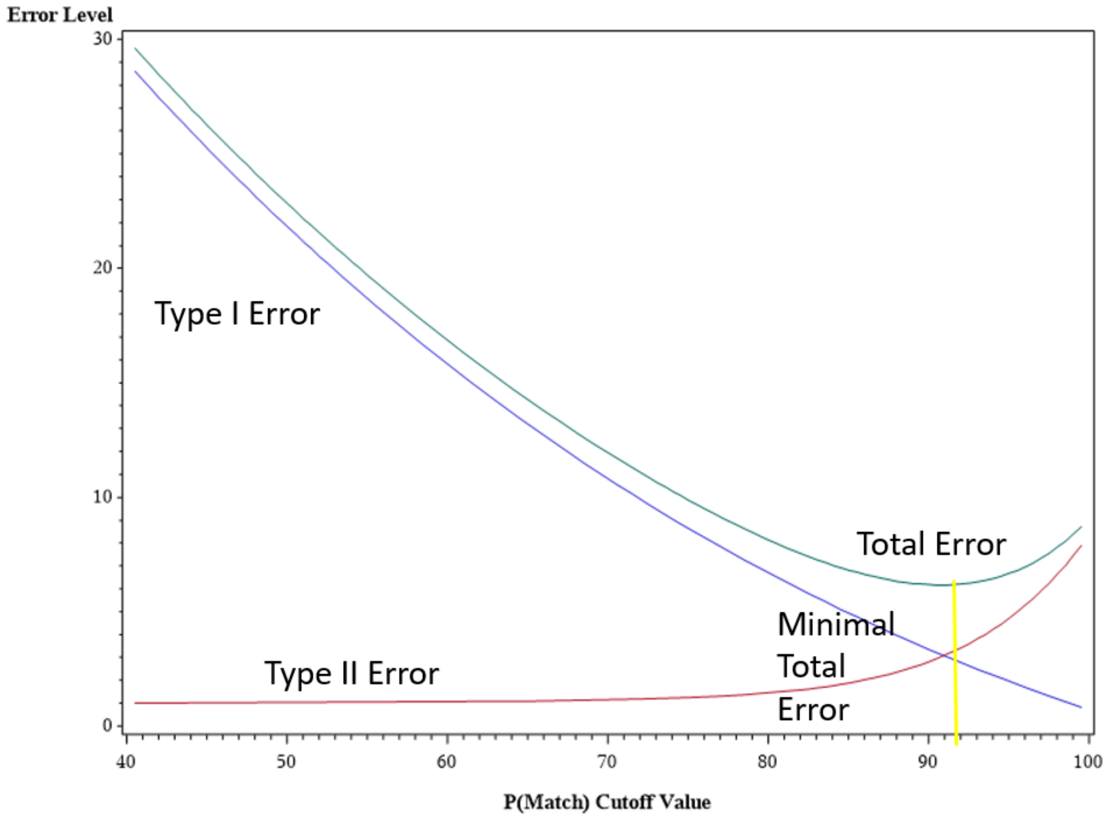
Because all records were included in the probabilistic linkage (i.e., even deterministic links), SSN agreement status (defined as 7 or more matching digits) was used to measure Type I error. Type I error for probabilistic links was measured as the total number of probabilistic links with non-agreeing SSN divided by the total number of probabilistic links with SSN available on both the survey and administrative record. Also, deterministically established links were considered to have 0% Type I error rates. While it was believed that the error for these links was quite small and near 0, it is expected that some error does exist even with the deterministically established links and so the estimate was likely biased low. Since a sizeable proportion of links were derived from the deterministic method, this had the effect of reducing the estimated Type I error by the proportion of probabilistically determined linkages among all linkages. For example, if the Type I error rate was estimated for probabilistic links as 1.2%, but only 40% of all links were derived from probabilistic analysis. Thus, for this example the estimated Type I error rate for the combined linkage process would be (0.40*0.012) = 0.0048 or 0.48%.

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

4.2 Set Probability Cutoff

One goal of record linkage is to have the lowest errors possible. However, as more pairs were accepted, pairs that were less certain to be matches as links increase the Type I error and decrease Type II error (see Figure 1). And as fewer pairs were accepted, pairs that were more certain to be matches as links decrease the Type I error and increase Type II error. The optimal trade-off is between Type I error and Type II error was not known, and likely this depends on the type of analysis to be conducted with the linked data, but it is assumed that it is not far from optimality when the sum of Type I and Type II error is at a minimum. For this reason, Type I and Type II error are estimated at various probability cut points and the one that showed the lowest estimate of total error was selected. For this linkage, the probability cutoff was set to 0.9225.

Figure 1: Error Level by Cutoff Value
(Schematic: not based on actual analysis)



4.3 Select Links Using Probability Threshold

The final step in the linkage algorithm was to determine links, which were pairs imputed to be matches. Links were pairs where the $Probvalid_{SSN_{Adj}}$ exceeded the set probability threshold (from Section 3.2). All pairs with an adjusted probability that fell below the set probability threshold were not linked.

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

4.4 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 1999-2018 NHIS and 1999-2018 NHANES linked HUD files, 24.0% and 26.2% of patients respectively, 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 0.9225 cutoff value, all HUD ID matches were assumed to represent the same individual. In the 1999-2018 NHIS and 1999-2018 NHANES, 4.6% and 4.9% of survey participants respectively, 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 0.09% and 2.06%, respectively.

Appendix II: Evaluation of NCHS–HUD Data Linkage: Comparison with HUD Program Recipients

This analysis compares values for the linked National Center for Health Statistics (NCHS) survey participants and all recipients of U.S. Department of Housing and Urban Development (HUD)-assisted housing. The percent distribution of selected characteristics from all HUD housing assistance recipients are presented side by side with that from the linked NCHS–HUD participants who received assistance during the same time period. This evaluation compares one year of data (2018) from the National Health Interview Survey (NHIS) and two cycles (2015–2018) of the National Health and Nutrition Examination Survey (NHANES) that have been linked with HUD data to the overall HUD administrative records for the corresponding time periods.

A similar comparison of linked and administrative data was conducted on a previous linkage (Vital and Health Statistics Reports Series 1, Number 60, October 2017 (cdc.gov)). However, the comparison was repeated with the new linked data for this report because this linkage relied on a new linkage methodology. The most significant difference between this linkage and the previous linkage is that the previous linkage was limited to deterministic linkage methods, whereas the current linkage used both deterministic and probabilistic methods. Also, for this linkage, the deterministic component was based on two levels of available SSN information: last 4-digit Social Security Number (SSN4) collected in the 2018 NHIS and 9-digit Social Security Number (SSN9) collected in all years of NHANES. The probabilistic linkage included participants that did not report an SSN.

The NCHS–HUD linked samples from the 2018 NHIS and the 2015–2018 NHANES are included in this analysis; however, the analysis compares HUD administrative data for the linked NCHS-HUD records to the overall HUD records. The HUD administrative data were used since they are the only data source available for both the NCHS– HUD linked sample and the population of HUD housing assistance recipients not included in the linked sample. Although some of these variables (e.g., sex, age, and total number of household members) are also collected in the NCHS surveys, the comparison relied on variables from the HUD administrative records. Therefore, there may be differences between the survey reported variables and the administrative variables that is not accounted for in this analysis. Analyses of these same variables obtained from the survey data may yield different results. This analysis is not intended to evaluate the HUD administrative data, but rather to present characteristics of both the NCHS–HUD linked sample and the population of HUD housing assistance recipients. The purpose of this analysis is to evaluate how the results compare for the linked NCHS-HUD sample and the overall HUD population. The following variables are included in the analysis:

- HUD program category — Housing Choice Voucher (HCV), multifamily (MF), public housing (PH), and other MF
- Total household income (<$10,000, $10,000 - $24,999, $25,000+ (for NHANES) and $25,000-49,999, $50,000+ (for NHIS))
- Age of the household member (0-17, 18-29, 30-44, 45-61, and 62 and over; age groups are based on categories used in HUD reports or publications)
- Indicator for households with at least one disabled person, based on HUD's operational definition of disability: https://portal.hud.gov/hudportal/ HUD?src=/program_offices/ fair_housing_equal_opp/disabilities/ inhousing
- Race/Ethnicity as documented in the HUD administrative record (White alone, non-Hispanic; Black alone, non-Hispanic; Asian alone, non-Hispanic; Other, non-Hispanic; Hispanic, any race)
- Total household members (1, 2, 3, 4, 5, and 6 or more (for NHANES), and 7 or more (for NHIS))

The following criteria are used for comparisons in the selected sociodemographic and household characteristics between the linked survey participants and all recipients of HUD-assisted housing:

- If the HUD population percentage fell inside the 95% Korn-Graubard adjusted Confidence Interval (CI) of the weighted estimate of the linked population, concurrent linked survey participants are described as "consistent or similar" to all recipients of HUD-assisted housing with respect to that characteristic.
- If the HUD population percentage fell outside the 95% Korn-Graubard adjusted CI of the weighted estimate of the linked population, then concurrent linked survey participants are described as "different (higher/more or lower/ less)" than all recipients of HUD-assisted housing with respect to that characteristic.

Because the two populations are not statistically independent and measures of correlation between the two populations were not readily available, comparisons were not statistically tested.

Evaluation of Linked 2018 NHIS–HUD Data: Methods

Recipients of HUD housing assistance during the 2018 calendar year were identified from HUD administrative records with effective dates of January 1, 2016 (for MTW) and July 1, 2017 (for all others) through December 31, 2018. Only the most recent HUD transaction was retained so that each individual was counted only once. In this analysis, these HUD housing assistance recipients are referred to as the population of "all 2018 HUD-assisted recipients". Concurrent 2018 NHIS–HUD participants are, in theory, a subset of this population. Similarities and differences between the concurrent 2018 NHIS– HUD participants and all 2018 HUD-assisted recipients demonstrate ways in which the linked data may or may not be comparable to the population of HUD housing assistance recipients. The eligibility-adjusted NHIS sample adult and NHIS sample child weights were used in this analysis to account for unequal probabilities of survey selection, nonresponse, and linkage eligibility. The CIs incorporate the complex sample design of NHIS.

Evaluation of Linked 2018 NHIS–HUD Data: Results

A total of 1,256 NHIS 2018 participants were concurrently linked to HUD administrative data. During the 2018 timeframe, the HUD administrative records contained data on 9,531,529 HUD recipients. Results are presented in Table 1.

For both populations, the HCV program had the highest participation (55.3% among all 2018 HUD-assisted recipients and 52.3% (CI: 46.2%, 58.2%) among concurrent 2018 NHIS–HUD participants). The percentage of all 2018 HUD-assisted recipients in PBS8 units was lower (21.5%) than among concurrent 2018 NHIS–HUD participants (27.2%, CI: 21.6%, 33.4%). All other types of HUD assistance were consistent between both populations.

The income distributions were consistent between both populations. The age distributions were also similar although the percentage of persons aged 18-29 was lower among all HUD-assisted recipients (9.7%) than the concurrent 2018 NHIS-HUD participants (14.8%, CI: 11.8%, 18.1%). The distribution of households with a disabled person and race/ethnicity distributions were similar between both populations.

Distribution of total household members was similar between the two populations in all categories except for the single member household category; the population of all 2018 HUD-assisted recipients had less individuals living in households with only 1 member (25.0%) than the population of concurrent 2018 NHIS–HUD participants (29.4%, CI: 25.3%, 33.7%).

Table 1. Number and percent distribution of concurrently linked 2018 NHIS–HUD participants and 2018 HUD-assisted recipients, by HUD program and selected HUD-collected demographic characteristics

	2018 HUD-assisted recipients		**2018 concurrently linked NHIS-HUD participants (*n* = 1,256)**			
					95% CI	
HUD program and housing Characteristics	***N***	**Percent**	***N***	**Percent (weighted)**	**Low**	**High**
HUD Program						
Public Housing	1,983,918	20.8	240	17.8	13.1	23.4
Project-Based Section 8	2,049,601	21.5	367	27.2	21.6	33.4
Housing Choice Voucher	5,272,169	55.3	602	52.3	46.2	58.2
Other	225,841	2.4	47	2.7	1.5	4.4
Total Annual Household Income						
<$10,000	3,569,174	37.5	499	39.0	35.1	43.2

$10,000-$24,999	4,179,525	43.9	588	41.4	37.4	45.5
$25,000-$49,999	1,547,522	16.2	147	16.2	12.8	20.0
$50,000-$99,000+	235,296	2.5	22	3.3	1.7	5.8
Age (years)						
0-17	3,464,876	36.4	312	33.0	29.6	36.6
18-29	926,885	9.7	122	14.8	11.8	18.1
30-44	1,776,477	18.6	199	16.0	13.6	18.6
45-61	1,493,167	15.7	242	16.4	13.8	19.2
62+	1,869,761	19.6	381	19.9	16.6	23.6
Disability Status						
Living with a Disability	2,117,555	22.2	379	23.8	20.6	27.3
Not Living with a Disability	7,413,974	77.8	877	76.2	72.7	79.4
Race/Ethnicity						
White alone, non-Hispanic	2,644,331	27.7	450	27.1	22.8	31.9
Black alone, non-Hispanic	4,419,185	46.4	479	43.4	38.0	48.8
Asian alone, non-Hispanic	274,809	2.9	27	2.4	1.3	4.0
Other, non-Hispanic	388,233	4.1	76	5.6	3.3	8.7
Hispanic (any race)	1,804,971	18.9	224	21.5	16.5	27.3
Total household members						
1	2,373,458	25.0	551	29.4	25.3	33.7
2	1,803,940	19.0	249	18.0	15.2	21.2
3	1,824,451	19.2	181	17.6	14.0	21.7
4	1,566,239	16.5	146	17.1	13.6	21.2
5	979,725	10.3	62	8.4	5.5	11.7
6	495,133	5.2	29	3.8	2.1	6.3
7 or more	455,837	5.0	15	*3.1	1.1	6.6

* Figure may be statistically unreliable because the absolute width of its Korn-Graubard confidence interval (CI) is between 5 and 30 percentage points, yet its relative width is larger than 130%.

NOTES: Confidence intervals (CIs) for percentages are calculated using the Korn-Graubard adjustment to the Clopper-Pearson method. NHIS is National Health Interview Survey, and HUD is U.S. Department of Housing and Urban Development.

SOURCES: NCHS, linked NHIS–HUD data sample, 2018; HUD administrative data, 2018.

Evaluation of Linked 2015-2018 NHANES–HUD Data: Methods

Two NHANES cycles were pooled due to the small sample size of concurrently linked NHANES–HUD participants for a single cycle. Recipients of HUD housing assistance with effective dates from January 1, 2014 (for MTW) and July 1, 2014 (for all others) through December 31, 2018 were identified from the HUD administrative records and included in the analysis to conservatively capture all HUD housing assistance recipients from the 2015–2018 calendar years. If more than one record existed per individual, only the most recent record was retained, so that each person was counted once at most. In this analysis, these HUD housing assistance recipients are referred to as the population of "all 2015–2018 HUD-assisted recipients." Concurrent 2015–2018 NHANES–HUD participants are, in theory, a subset of this population. Similarities and differences between the concurrent 2015–2018 NHANES–HUD participants and all 2015–2018 HUD-assisted recipients demonstrate ways in which the linked data may or may not be comparable to the population of HUD housing assistance recipients.

Linkage eligibility-adjusted sample weights were used in NHANES comparison analysis to account for unequal probabilities of survey selection, survey nonresponse, and linkage eligibility. The confidence intervals reflect the complex design of NHANES.

Evaluation of Linked 2015-2018 NHANES–HUD Data: Results

A total of 798 NHANES 2015–2018 participants were concurrently linked to HUD administrative data. During the 2015–2018 timeframe, the HUD administrative records contained data on 13,645,177 HUD recipients. Results are presented in Table 2.

In both populations, the HCV program was the most common housing assistance program. HUD program utilization was consistent between both populations.

The income distributions were also consistent between both populations. The age distributions were similar although the percentage of persons aged 0-17 was higher (40.2%) in the overall HUD-assisted population than the concurrent 2015-2018 NHANES-HUD participants (35.1%, CI: 30.4%, 40.0%).

The distribution of households with a disabled person was similar between both populations. The race/ethnicity distributions differed among persons who identified as Asian alone, non-Hispanic. Among all 2015–2018 HUD-assisted recipients, the percentage of Asian, non-Hispanic persons was lower (2.6%) than concurrent 2015–2018 NHANES–HUD participants (5.0%, CI: 2.7%, 8.4%).

Distribution of total household members differed between the two populations in households with 6 or more members; the population of all 2015–2018 HUD-assisted recipients had more individuals living in households with 6 or more members (9.9%) than the population of concurrent 2015–2018 NHANES–HUD participants (6.6%, CI:4.6%, 9.1%)

Table 2. Number and percent distribution of concurrently linked 2015–2018 NHANES–HUD participants and 2015–2018 HUD-assisted recipients, by HUD program and selected HUD-collected demographic characteristics

	2015-2018 HUD-assisted recipients		2015-2018 concurrently linked NHANES-HUD participants (*n* = 798)			
					95% CI	
HUD program and housing Characteristics	***N***	**Percent**	***N***	**Percent (weighted)**	**Low**	**High**
HUD Program						
Public Housing	3,171,956	23.3	295	22.1	17.0	27.9
Project-Based Section 8	3,048,798	22.3	322	23.2	16.9	30.5
Housing Choice Voucher	7,060,255	51.7	641	48.3	38.8	57.8
Other	364,168	2.7	86	*6.5	2.3	14.0
Total Annual Household Income						
<$10,000	5,667,212	41.5	551	42.1	37.1	47.2
$10,000-$24,999	5,993,178	43.9	593	43.7	38.4	49.1
$25,000+	1,984,781	14.6	200	14.3	10.3	19.0

Age (years)						
0-17	5,483,624	40.2	642	35.1	30.4	40.0
18-29	1,474,431	10.8	119	14.7	10.7	19.3
30-44	2,514,885	18.4	140	15.8	13.0	18.9
45-61	1,994,416	14.6	167	15.0	11.8	18.7
62+	2,177,565	16.0	276	19.5	13.5	26.8
Disability Status						
Living with a Disability	2,604,940	19.1	225	18.0	14.9	21.5
Not Living with a Disability	11,040,237	80.9	1,119	82.0	78.5	85.1
Race/Ethnicity						
White alone, non-Hispanic	3,788,852	27.8	184	23.3	16.7	31.1
Black alone, non-Hispanic	6,238,768	45.7	807	54.3	44.6	63.7
Asian alone, non-Hispanic	356,851	2.6	85	5.0	2.7	8.4
Other, non-Hispanic	576,741	4.2	57	3.4	1.8	5.8
Hispanic (any race)	2,683,965	19.7	211	14.0	8.2	21.6
Total household members						
1	3,058,078	22.5	291	24.5	18.5	31.3
2	2,639,491	19.4	231	19.2	15.8	22.9
3	2,764,493	20.3	255	18.9	15.3	22.8
4	2,352,368	17.3	269	18.7	15.0	22.8
5	1,446,365	10.6	181	12.2	8.9	16.3
6 or more	1,346,139	9.9	113	6.6	4.6	9.1

* Figure may be statistically unreliable because the absolute width of its Korn-Graubard confidence interval (CI) is between 5 and 30 percentage points, yet its relative width is larger than 130%.

NOTES: Confidence intervals (CIs) for percentages are calculated using the Korn-Graubard adjustment to the Clopper-Pearson method. NHANES is National Health and Nutrition Examination Survey, and HUD is U.S. Department of Housing and Urban Development.

SOURCES: NCHS, linked NHANES–HUD data sample, 2015–2018; HUD administrative data, 2015–2018.

Conclusion

The linked NCHS–HUD data appears to be generally comparable to the overall population of HUD recipients with the noted exceptions.

Among concurrent linked NHIS–HUD and NHANES–HUD participants, the distributions of income and disability status were similar to the corresponding percentage among all HUD recipients. Differences were observed by program category of HUD housing assistance received, age, race/ethnicity, and total number of household members. Differences in characteristics that were not measured here between concurrent linked survey participants and the population of HUD housing assistance recipients should also be taken into consideration.

These differences may impact the generalizability of study results from analyses of the linked NCHS–HUD data to the overall population of HUD housing assistance recipients.

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

The restricted-use NCHS-linked files are merged with the public-use NCHS survey data files using unique person identifiers. Therefore, it is important for researchers to include the correct survey person identification number: PUBLICID (for NHIS), or SEQN (for NHANES).

For using NHIS data, it also is important to note in the descriptions below that the variable names and locations needed to construct PUBLICID vary by NHIS survey year

Note: At this time the linked HUD data files are only available for research use through the NCHS restricted access data center (RDC). Approved RDC researchers may choose to provide their own analytic files created from public use survey files to the RDC. Therefore, it is important for researchers to include a survey-specific Public Identification number on any analytic files sent to the RDC. The RDC will merge data (using PUBLICID, SEQN or RESNUM) from the linked 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, SEQN or RESNUM is provided below.

NHIS 1999-2003

The data items 'Survey year' (SRVY_YR), 'Household number' (HHX), 'Family number' (FMX), and 'Person number' (PX) identify a participant within each NHIS*. These data items must be concatenated to obtain the unique personal identifier (PUBLICID) used in the NHIS-HUD linked file.

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

SAS example:

```
length publicid $14;

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

Note: The SAS input statements available from the NHIS public-use data website do NOT input all of the variables as character and they must be in character format for the concatenation.

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

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

*The data item 'Person number' was called PX in the 1999-2003 NHIS and FPX in the 2004-2012 NHIS. Users may find it necessary to create an FPX variable in the 2003 and earlier datasets (or PX in later datasets).

NHIS 2004

Taken together, the data items ‘Survey year’ (SRVY_YR), ‘Household number’ (HHX), ‘Family number’ (FMX), and ‘Person number’ (FPX) identify a participant within NHIS 2004. These data items must be concatenated to obtain the unique personal identifier (PUBLICID) used in the NHIS-HUD linked file.

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

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)
```

NHIS 2005 – 2018

Taken together, the data items ‘Survey year’ (SRVY_YR), ‘Household number’ (HHX), ‘Family number’ (FMX), and ‘Person number’ (FPX) identify a participant within each NHIS. These data items must be concatenated to obtain the unique personal identifier (PUBLICID) used in the NHIS-HUD linked file.

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

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)
```

Appendix IV: 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 in 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);
```