

The Linkage of the 2019 and 2020 National Hospital Care Survey (NHCS) to 2018-2020 and 2019-2021 U.S. Department of Housing and Urban Development Administrative Data: Linkage Methodology and Analytic Considerations

Data Release Date: August 25, 2025

Document Version Date: August 25, 2025

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 2019 and 2020 National Hospital Care Survey to 2018-2020 and 2019-2021 U.S. Department of Housing and Urban Development Administrative Data: Linkage Methodology and Analytic Considerations*, August 2025. Hyattsville, Maryland. Available at the following address: <https://www.cdc.gov/nchs/linked-data/nhcs/restricted-hud.html>

Contents

List of Acronyms.....	5
1 Introduction.....	6
2 Background on Linked Files.....	6
2.1 National Hospital Care Survey.....	6
2.2 U.S. Department of Housing and Urban Development (HUD) Public and Assisted Housing Programs and Data.....	7
2.2.1 HUD Public and Assisted Housing Programs.....	7
2.2.2 HUD Administrative Data.....	8
3 Linkage Methodology.....	9
3.1 Linkage Eligibility Determination.....	9
3.2 Overview of Linkage.....	9
4 Analytic Considerations.....	12
4.1 Analytic Considerations for NHCS Data.....	12
4.1.1 NHCS Restricted-Use Files (RUF).....	12
4.1.2 NHCS Hospital Eligibility and Sampling.....	12
4.1.3 2020 NHCS Encounter Weights.....	12
4.1.4 NHCS Patient Identification Number.....	13
4.2 Analytic Considerations for Linked HUD Data Files.....	13
4.2.1 Description of Linked NHCS - HUD Data Files.....	13
4.2.1.1 HUD Match File.....	13
4.2.1.2 Linked HUD Program Participation Files.....	14
4.2.2 Identification of Ever and Concurrent HUD-Assisted Patients.....	14
4.2.2.1 Ever Received HUD-assisted Housing.....	14
4.2.2.2 Temporal Alignment of HUD Assistance with NHCS Patient Encounter Data.....	14
4.2.2.3 Temporal Alignment of HUD Assistance with NHCS Patient Data.....	16
5 Access to Data Files.....	17
5.1 Access to the Restricted-Use Linked NHCS – HUD Administrative Data Files.....	17
5.2 Merging NHCS Survey Data to the Linked NHCS-HUD Administrative Data Files.....	17
5.3 Additional Related Data Sources-Linked NHCS-NDI Mortality Files.....	17
Appendix I: Detailed Description of Linkage Methodology.....	18
1 NHCS and HUD Linkage Submission Files.....	18
2 Deterministic Linkage Using Unique Identifiers.....	20
3 Probabilistic Linkage.....	20

3.1 Blocking.....	21
3.2 Score Pairs	22
3.2.1 M and U Probabilities.....	23
3.2.2 M and U Probabilities for First and Last Names	25
3.2.3 Adjustment of U-Probabilities for Alternate Submission Records.....	26
3.2.4 Calculate Agreement and Non-Agreement Weights	29
3.2.5 Calculate Pair Weight Scores.....	29
3.3 Probability Modeling.....	30
3.4 Adjustment for SSN Agreement	31
4 Estimate Linkage Error, Set Probability Cut-off Value, and Select Matches	32
4.1 Estimating Linkage Error to Determine Probability Cut-off Value	32
4.2 Set Probability Cut-off Value.....	33
4.3 Select Links Using Probability Cut-off Value	34
4.4 Resolving NHCS Patient IDs that Linked to Multiple HUD Records.....	34
4.5 Computed Error Rates of Selected Links.....	35

List of Acronyms

CCD, Continuity of Care Document
DHCS, Division of Health Care Statistics
DOB, date of birth
ED, emergency department
EHR, electronic health record
E-M, expectation-maximization
ERB, Ethics Review Board
HCV, Housing Choice Voucher program
IP, inpatient
MF, Multi-family housing programs
MTW, Moving to Work demonstration program
NCHS, National Center for Health Statistics
NDI, National Death Index
NHCS, National Hospital Care Survey
OP, outpatient
OPD, outpatient department
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
SSN, Social Security number
TRACS, Tenant Rental Assistance Certification System
UB-04, uniform billing form

1 Introduction

As the nation's principal health statistics agency, the mission of the National Center for Health Statistics (NCHS) is to provide statistical information that can be used to guide actions and policy to improve the health of the American people. In addition to collecting and disseminating the Nation's official vital statistics, NCHS conducts several population-based surveys and healthcare establishment surveys, including the National Hospital Care Survey (NHCS) <https://www.cdc.gov/nchs/nhcs/index.html> (accessed July 2025). The 2019 NHCS collected data from a sample of 598 hospitals of which 82 provided linkage-eligible patient data and the 2020 NHCS collected data from a sample of 608 hospitals of which 106 provided linkage-eligible patient data. For participating hospitals, these data cover all hospital encounters to the inpatient and emergency department occurring throughout the calendar year. The NHCS includes detailed information about hospital characteristics, patients' characteristics, and treatment. Even though NHCS is an establishment survey (i.e., hospitals are the sampling unit) it collects patient personally identifiable information (PII), which enable data linkages.

Through its data linkage program, NCHS has been able to expand the analytic utility of the data collected from NHCS by augmenting it with housing assistance program data collected by the U.S. Department of Housing and Urban Development (HUD). This report will describe the linkage of the 2019 NHCS to 2018-2020 HUD administrative data and the 2020 NHCS to 2019-2021 HUD administrative data. The linkage of NHCS patient data with HUD administrative housing assistance program participation data creates a new data resource that can support a wide range of public health surveillance and policy evaluation studies focused on the relationship between housing and health.

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

The data linkage work was performed at NCHS through contract #HHS75D30123A17667 by NORC at the University of Chicago with funding from the Department of Health and Human Services' Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF).

2 Background on Linked Files

2.1 National Hospital Care Survey

The NHCS is an establishment survey that collects inpatient (IP), emergency department (ED) episode-level data from sampled hospitals. It is one of the National Health Care Surveys, a family of surveys covering a broad spectrum of healthcare delivery settings including ambulatory, hospital, and post-acute and long-term care providers. The goal of the NHCS is to provide reliable and timely healthcare utilization data for hospital-based settings, including prevalence of conditions, health status of patients, health services utilization, and substance-involved ED visits.

From participating hospitals, NHCS collects data on all IP and ED visits occurring during the calendar year. In previous years of the survey, hospitals were required to provide data from claims records, but to reduce the burden of reporting on participating hospitals, for the 2019 and 2020 data collection hospitals were given the option of providing their data in the form of either Uniform Billing (UB)-04 administrative claims records or electronic health records (EHR). Additionally, data could be provided by third-party entities, like Vizient or (starting in 2020) the American College of Emergency Physicians. Participating

hospitals were required to submit one type of data (e.g., UB-04 administrative claims or EHR, not both). For those hospitals submitting EHR data, this was submitted in the format of HL7 CDA® R2 Implementation Guide: National Health Care Surveys Release 1, DSTU Release 1.2 – US Realm (http://www.hl7.org/implement/standards/product_brief.cfm?product_id=385). NHCS collects patient PII (e.g., name, date of birth, and Social Security Number (SSN)), which allows for the linkage of episodes of care across hospital units as well as to other data sources, such as HUD data. The linkage described here includes only IP and ED visits.

2.2 U.S. Department of Housing and Urban Development (HUD) Public and Assisted Housing Programs and Data

2.2.1 HUD Public and Assisted Housing Programs

The U.S. Department of Housing and Urban Development (HUD) is the federal agency responsible for overseeing domestic housing programs and policies. While HUD is responsible for administering various housing and community development programs, the linkage with the 2019 and 2020 NHCS focuses on HUD's three largest housing assistance programs: Housing Choice Vouchers, Public Housing, and assisted Multifamily programs. Persons and households participating in these program types are referred to as "HUD-assisted" in this document.

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

The HUD Housing Choice Voucher (HCV) program is the federal government's largest housing assistance program, allowing families with lower incomes, older adults (persons 62 or older), and persons with disabilities to choose and lease safe and affordable housing. In the HCV program, housing assistance is tenant-based, meaning participants find their own housing in the private market. Participants are free to choose any housing unit that meets program requirements. In the NHCS-HUD linked data, the HCV program also includes several smaller programs including: Homeownership Vouchers, Project-Based Vouchers, Section 8 Moderate Rehabilitation, and the Section 8 Rental Certificate programs. Among the 2019 and 2020 NHCS patients that linked to HUD, approximately 56% (2019 NHCS) and 51% (2020 NHCS) were participating in an HCV program.

The public housing (PH) program was established to provide safe rental housing for eligible low-income families, the elderly, and persons with disabilities. HUD provides capital subsidies and operating subsidies to local Public Housing Agencies (PHAs) that manage public housing for eligible low-income residents. Unlike the HCV program, PH is project-based meaning tenants do not choose their housing but are instead assigned housing in a specific unit, building, or development. Approximately 20% of 2019 NHCS patients and 24% of 2020 NHCS patients linked to HUD were participating in a PH program.

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

development of the linked NHCS - HUD administrative data files. (See [Section 2.2.2](#) for more information on HUD administrative data).

The assisted multi-family housing (MF) program category in the linked NHCS – HUD data encompasses a number of separate, distinct HUD programs, including: Project-Based Section 8 (PBS8) the largest MF program, Section 221(d)(3) Below Market Interest Rate, Section 236 Multi-family 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. Approximately 28% of 2019 NHCS patients and 29% of 2020 NHCS patients that linked to HUD were participating in any MF program.

2.2.2 HUD Administrative Data

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

PIC contains household-level and person-level administrative records pertaining to persons and households participating in HUD’s HCV and PH program types. The underlying forms used to capture information for these programs are the [HUD-50058](#) and the [HUD-50058MTW](#). The PIC data extract created for the NHCS linkage with HUD data 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 (March, June, September, and December).

TRACS is a system developed to collect and maintain certified tenant data from owners and management agents of MF housing programs. The underlying form used to capture information for MF programs is [HUD-50059](#). The TRACS data extract created for the NHCS-HUD data linkage was based on TRACS point-in-time quarterly extracts from the TRACS production system. Similar to the PIC data, 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 and monthly HUD program participation variables for the final NHCS HC-HUD linked data files. For more detailed information on the specific HUD data available on the NHCS-HUD linked data files, see [Section 4.2.1](#).

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 July 2025).

3 Linkage Methodology

3.1 Linkage Eligibility Determination

The linkage of 2019 and 2020 patient records to HUD data was conducted through a designated agent agreement between NCHS and HUD. Approval for the linkage was provided by NCHS's Research Ethics Review Board (ERB).¹

Linkage was attempted only for NHCS patient records that had at least two of the following three identifiers present:

- valid SSN²
- valid date of birth (month, day, and year)³
- valid name (first, middle initial, and last)⁴

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

The variable ELIGSTAT, included on the linked 2019 and 2020 NHCS-HUD match files, provides the linkage eligibility status for each NHCS patient record. ELIGSTAT values include 0 (ineligible) or 1 (eligible). The 2019 and 2020 NHCS included 2,744,837 (57.5%) and 3,117,460 (57%) patients, respectively, who were determined to be eligible for linkage with HUD administrative data. Note that linkage eligibility is distinct from program eligibility, which defines whether a person meets the eligibility criteria for a specific government-administered or funded program.

3.2 Overview of Linkage

This section outlines steps used to link the NHCS data to HUD enrollment data. The linkage was conducted at a patient level using patient identifiers collected from hospital submitted patient encounter records. The patient identifiers collected from the encounter records were used to link HUD program participation records covering the calendar year one year prior to the NHCS survey year through one calendar year after the survey year. Therefore, the HUD linkage period for the 2019 NHCS includes HUD program participation data from 2018 through 2020 and the 2020 NHCS includes HUD program participation data from 2019 through 2021. For more detailed information on linkage methodology see [Appendix I: Detailed Description of Linkage Methodology](#).

Linkage-eligible NHCS patient records 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 NHCS patient 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-

¹ The NCHS ERB is an appointed ethics review committee that is established to protect the rights and welfare of human research subjects.

² Nine-digit SSN is considered valid if: 9-digits in length, containing only numbers, does not begin with 000, 666, or any values after 899, all 9-digits cannot be the same (i.e., 111111111, etc.), middle two and last 4-digits cannot be 0's (i.e., xxx-00-xxxx or xxx-xx-0000), and digits are not consecutive (ex. 012345678). Additionally, special SSN values (i.e., 123-123-1234, 111-22-3333, 010-010-0101, 001-01-0001, etc.) were changed to missing. Four-digit SSN is considered valid if: 4-digits in length, containing only numbers, and is between 0001 and 9999.

³ A date of birth is considered valid if at least two of the three date parts are valid date values.

⁴ A name is considered valid if: either first or last name has two or more characters, and two of the three name parts (first, middle initial, and last) are non-missing.

Sunter method.⁵ Following this, a selection process was implemented with the goal of selecting pairs that represented the same individual between the two data sources.

It is important to note that both deterministic and probabilistic linkages were conducted separately for each sex category (males and females). That is, records on the NHCS and HUD linkage submission files were first separated using the recorded sex value and then each set of files were separately linked.⁶ The Fellegi-Sunter method assumes independence between the agreement status of variables used to score the records. Because first names are commonly associated with sex, running the linkage separately by sex ensures independence and enables more appropriate weighting of name comparisons when using the methods described by Fellegi-Sunter.⁷ The following steps were implemented:

1. Separate NHCS and HUD submission files using sex. HUD submission records were further restricted to the combined linkage period for both of the NHCS surveys (i.e., 2018-2021).
2. Deterministic linkage joined records on exact SSN and validated links by comparing other identifying fields (i.e., first name, last name, day of birth, etc.)
3. Probabilistic linkage identified likely matches, or links, between all records. All records were probabilistically linked and scored as follows:
 - a. Formed pairs via blocking
 - b. Scored pairs
 - c. Modeled probability - assigned estimated probability that pairs are links
4. Pairs were selected that were believed to represent the same individual between data sources (i.e., they are a match).
 - a. Deterministic matches (from step 2) were assigned a match probability of 1
 - b. Record pairs selected from the probabilistic match (step 3) were assigned the model match probability. Record pairs with a match probability above the established probability cut-off value were determined to be matches.

For each NHCS patient record that was deemed a match, HUD extracted information from the PICS and TRACS systems and sent them to NCHS through a secure data transfer system.

[Table 1](#) presents the total number of NHCS patients by age group and sex, the number who were eligible for linkage, the number who were linked to HUD administrative data, and the unweighted percentage of all patients and those eligible for linkage who were linked to HUD administrative program data.

⁵ Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," JASA 40 1183-1210.

⁶ Before the submission records were separated, alternate submission records (see [Appendix I, section 1](#)) with an imputed sex value of male and female were created for records with missing sex.

⁷ First names are often sex specific (i.e., first name Robert is usually associated with males and Mary is usually associated with females). Additionally, multiple part first and last names are more likely to be associated with females, which are handled differently when creating the linkage submission file. See [Table 2](#) in Appendix I, Section 1 for additional information on the alternate record generation process for multiple part names.

Table 1. 2019 NHCS linked to 2018-2020 HUD Administrative Records and 2020 NHCS linked to 2019-2021 HUD Administrative Records: Sample Sizes and Percent Linked, by Age and Sex

	Sample Size		Percent Linked		
	Total Sample	Eligible for Linkage ²	Linked to HUD Administrative Data ³	Total Sample ⁴	Eligible Sample ⁵
2019 NHCS					
Age¹					
0-17	732,617	732,524	59,728	8.15	8.15
18-44	879,557	879,466	59,917	6.81	6.81
45-64	572,101	572,057	34,762	6.08	6.08
65 and over	560,618	560,608	31,163	5.56	5.56
Not Calculated	2,030,271	182	6	0.00	3.30
Total	4,775,164	2,744,837	185,576	3.89	6.76
Sex					
Male	1,244,612	1,244,490	61,806	4.97	4.97
Female	1,499,944	1,499,828	123,743	8.25	8.25
Missing	2,030,608	519	27	0.00	5.20
Total	4,775,164	2,744,837	185,576	3.89	6.76
2020 NHCS					
Age¹					
0-17	634,432	634,411	49,806	7.85	7.85
18-44	1,091,235	1,091,216	75,171	6.89	6.89
45-64	705,806	705,792	47,006	6.66	6.66
65 and over	685,695	685,686	42,845	6.25	6.25
Not Calculated	2,351,218	355	7	0.00	1.97
Total	5,468,386	3,117,460	214,835	3.93	6.89
Sex					
Male	1,425,163	1,425,121	69,553	4.88	4.88
Female	1,691,786	1,691,749	145,233	8.58	8.58
Missing	2,351,437	590	49	0.00	8.31
Total	5,468,386	3,117,460	214,835	3.93	6.89

NOTES: Data are presented at patient level.

¹ Age is as of final encounter (date of last known contact). Age is calculated by subtracting patient date of birth (DOB) from the final encounter date. When more than one DOB was present, the minimum of the non-missing DOB was selected.

² Eligibility for linkage is based upon having sufficient PII in at least two of three data element groups: SSN, name, and date of birth.

³ This group includes linkage-eligible patients who linked to HUD enrollment database at any time during the linkage interval (2019 NHCS: 2018-2020 HUD, 2020 NHCS: 2019-2021).

⁴ This percentage is calculated by dividing the number of linked patients by the number of patients in the total sample.

⁵ This percentage is calculated by dividing the number of linked patients by the total number of linkage-eligible patients.

4 Analytic Considerations

This section summarizes some key analytic issues for users of the linked NHCS data and HUD administrative records. It is not an exhaustive list of the analytic issues that researchers may encounter while using the linked NHCS-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 NHCS-HUD linked data files are encouraged to read “[A Primer on HUD Programs and Associated Administrative Data](#)” (accessed July 2025) for additional information on HUD program and corresponding administrative data, including important analytic considerations.

4.1 Analytic Considerations for NHCS Data

4.1.1 NHCS Restricted-Use Files (RUF)

The 2019 and 2020 NHCS restricted-use survey data are made available for research use through the NCHS RDC network. For more information about obtaining access to NHCS RUFs see [Section 5](#). The NHCS RUFs are organized as relational data tables organized by Facility, Patient, Encounter, Conditions, Services, Current Procedural Terminology (CPT)/Healthcare Common Procedure Coding System (HCPCS) Flag, Revenue Code Flag, and Encounter Weights. For more information about the specific variables and the observational unit for each data table please see the NHCS data documentation available at: <https://www.cdc.gov/rdc/data/b1/NHCS-RDC-Data-Dictionary.pdf> (accessed July 2025)

4.1.2 NHCS Hospital Eligibility and Sampling

Eligible hospitals for NHCS are non-institutional, non-federal hospitals with six or more staffed IP beds. There were 6,906 hospitals which met these criteria as of 2020 to form the survey frame. A base sample of 500 hospitals and a reserve sample of 500 additional hospitals was drawn from this frame. Initially, the base sample of 500 hospitals was fielded. In 2017, the sample and frame files were updated to include newly constructed hospitals from a new source file. Updates to the NHCS sample and frame occur every three years. Due to the addition of newly sampled birth hospitals, the sample increased to 598 hospitals in 2019 and 608 hospitals in 2020. Moreover, of the 598 sampled hospitals in 2019, 82 hospitals were eligible for linkage and of the 608 sampled hospitals in 2020, 106 hospitals were eligible for linkage (note: the linkage-eligible number excludes hospitals that provided records covering less than 6 months of the analysis period). In 2019, of the 82 linkage-eligible hospitals, 82 hospitals sent IP data, and 76 hospitals sent ED data. Of the 106 linkage-eligible hospitals in 2020, 104 hospitals sent IP data, and 100 hospitals sent ED data. Although NHCS collected outpatient data from 2013-2016, outpatient data are no longer being collected.

4.1.3 2020 NHCS Encounter Weights

While national estimates of hospital encounters in the IP and ED are not available for 2019 NHCS due to low response rates, the 2020 NHCS can produce nationally representative estimates. The Division of Health Care Statistics (DHCS) at NCHS has produced encounter level weights that can be used for national estimates of hospital encounters in the IP and ED. For information regarding producing weighted estimates with 2020 NHCS data the NHCS data documentation available at: <https://www.cdc.gov/nchs/data/nhcs/2020-NHCS-PUF-Tech-Doc-508.pdf> (accessed July 2025)

The linkage between the NHCS data and HUD housing assistance program data was conducted at a patient level using patient identifiers collected from patient encounter records. The patient identifiers collected from the encounter records were used to link HUD program participation records for NHCS patients. Although DHCS has developed encounter level weights for use with the 2020 NHCS encounter data, patient level weights for use with the linked HUD data are not available.

4.1.4 NHCS Patient Identification Number

Each patient in the NHCS is assigned a unique identification number, PATIENT_ID. PATIENT_ID does not contain any identifiable information about the patient and is intended to be unique for each individual receiving IP or ED services at a participating hospital. However, the de-duplication of patient records required to generate this ID depends on sometimes incomplete or erroneous data, there may be instances where the same individual is represented by more than one PATIENT_ID. This happens infrequently and should not greatly impact analyses.⁸

4.2 Analytic Considerations for Linked HUD Data Files

4.2.1 Description of Linked NHCS - HUD Data Files

4.2.1.1 HUD Match File

The linked HUD Match file can be used to identify which of the NHCS patients were eligible for linkage and linked to a HUD record. This file contains one record for each unique NHCS patient ID and contains the variables ELIGSTAT, PROBVALID, and HUD_MATCH_STATUS.

The variable ELIGSTAT should be used to determine linkage eligibility ([Section 3.1](#)). NHCS patient IDs with an ELIGSTAT value of 1 were considered eligible for linkage to the HUD enrollment records.

This file also contains information on the estimated probability of match validity (PROBVALID). An estimated probability of match validity was computed for each candidate pair and compared against a probability cut-off value to determine which pairs were links (an inferred match). For additional discussion on how PROBVALID was estimated, see Appendix I, Sections [3.3](#) and [3.4](#). NCHS used a probability cut-off value which minimized the total estimated counts of Type I error (false positive links – identified as participating in a HUD program but actually are not) and Type II error (false negative links – identified as not participating in a HUD program but actually are).

In the 2019 and 2020 NHCS – HUD linkages, NCHS used a probability cut-off value of 0.8925 to determine final match status. Candidate pairs with a PROBVALID that exceeded the probability cut-off value (i.e., PROBVALID>0.8925) were deemed a link. For additional discussion on probability cut-off value determination and record selection, please see [Appendix I, Section 4](#). For some analyses, it may be desirable to reduce the Type I error. To do this, researchers should increase the probability cut-off value to a value closer to 1.0. Researchers wishing to increase the probability cut-off value should request PROBVALID in their RDC proposal. Note, the probability cut-off value cannot be decreased from 0.8925 as pairs estimated with lower match probability are not made available to researchers.

The HUD_MATCH_STATUS variable can be used to identify which of the NHCS patients were participating in any HUD program during the HUD linkage period. When equal to one, HUD_MATCH_STATUS indicates that a NHCS patient was matched to at least one HUD housing assistance program administrative

⁸ For more information on Patient_ID generation, see Technical Notes on page 15: <https://www.cdc.gov/nchs/data/nhsr/nhsr097.pdf> (accessed June 2, 2025)

transaction record during the linkage period, either 2018-2020 for NHCS 2019 or 2019-2021 for NHCS 2020.

4.2.1.2 Linked HUD Program Participation Files

The NHCS data have been linked to multiple years of HUD data. HUD program participation data may be available for patients during the NHCS data collection year as well as the calendar year prior to or after the survey period.

The linked HUD program specific participation files contain monthly indicator variables to indicate whether a linked NHCS patient received HUD housing assistance benefits within a given month during the 3-year linkage period. There are four HUD program participation files, including a summary program participation file (any HUD program participation) and then three program specific participation files for each of the three main HUD housing assistance programs (HCV, PH, and MF).

Each of the HUD program-specific participation files contains one record for each linked NHCS patient ID and 36 monthly HUD participation indicators (one for each month during the linked data time span). The monthly indicators are created from program participation episodes that were derived using the transaction dates from the HUD transaction file. For each month in which at least one day of HUD participation is identified, the monthly indicator is set to 1, indicating program participation for that month. Monthly indicator variables for months with no HUD program participation are set to 0. For example, for NHCS 2020, a HUD program participation episode that began in June 2020 and ended in December 2021 would have monthly indicator variables set to 1 for all months from June 2020 through December 2021. All other monthly program participation variables from January 2019 to May 2020 would be set to 0. It is not possible for analysts using the HUD program-specific participation files to determine whether a HUD program participation period began prior to or ended after the NHCS - HUD linkage period (January 2019 – December 2021 for 2020 NHCS and January 2018 – December 2020 for 2019 NHCS).

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 July 2025).

4.2.2 Identification of Ever and Concurrent HUD-Assisted Patients

4.2.2.1 Ever Received HUD-assisted Housing

To identify NHCS patients who were participating in a HUD-assisted housing program at any time during the linkage period, researchers should use HUD_MATCH_STATUS in the linked HUD Match file. A HUD_MATCH_STATUS value of 1 indicates that an NHCS patient ID was linked with a HUD record at least once during the HUD linkage period (calendar years 2018-2020 for NHCS 2019 and 2019-2021 for NHCS 2020). To determine which months the NHCS patient was ever enrolled in a HUD assistance program, researchers should use the monthly program participation variables found in the Any HUD Program Participation File. To determine which of the specific HUD programs the patient was participating in, researchers should use the program specific participation files (HCV, MF and PH).

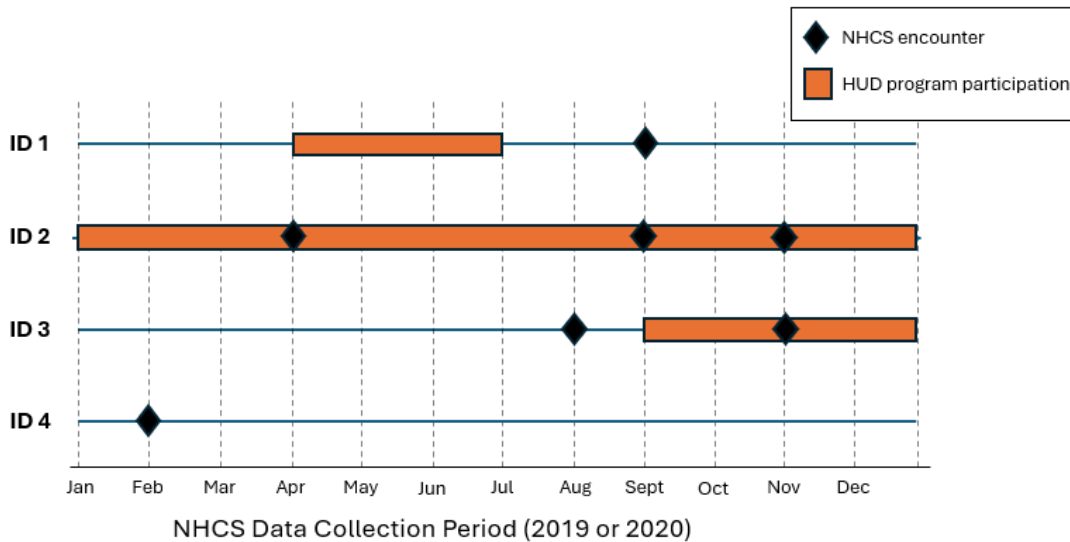
4.2.2.2 Temporal Alignment of HUD Assistance with NHCS Patient Encounter Data

To identify whether the NHCS patient was enrolled in a HUD program during, before, or after a specific hospital encounter, researchers can compare the month and year of the NHCS patient visit with the monthly indicators included in the linked HUD program participation data files (Any, HCV, MF, and PH).

[Figure 1](#) below depicts potential temporal alignment scenarios for NHCS patient encounter data and

monthly HUD program participation data for four hypothetical patients noted as patient ID1 through patient ID4. In each timeline, the diamond represents the month during which the NHCS patient encounter occurred, and the orange bar represents the month(s) during which the patient received HUD assistance. For example, Patient ID1 was participating in a HUD program during the months of April, May, and June but did not participate in July. The orange bar of Patient ID 1 spans only the months April through June and stops at July.

Figure 1. Temporal Alignment of NHCS Data Linked to HUD Administrative Data Files during NHCS Data Collection Period.



Notes: HUD is the U.S. Department of Housing and Urban Development
Sources: Hypothetical example based on NHCS patient data linked to HUD administrative data for data collection period.

The examples shown in [Figure 1](#) are as follows and have the same interpretation for 2019 and 2020 NHCS,

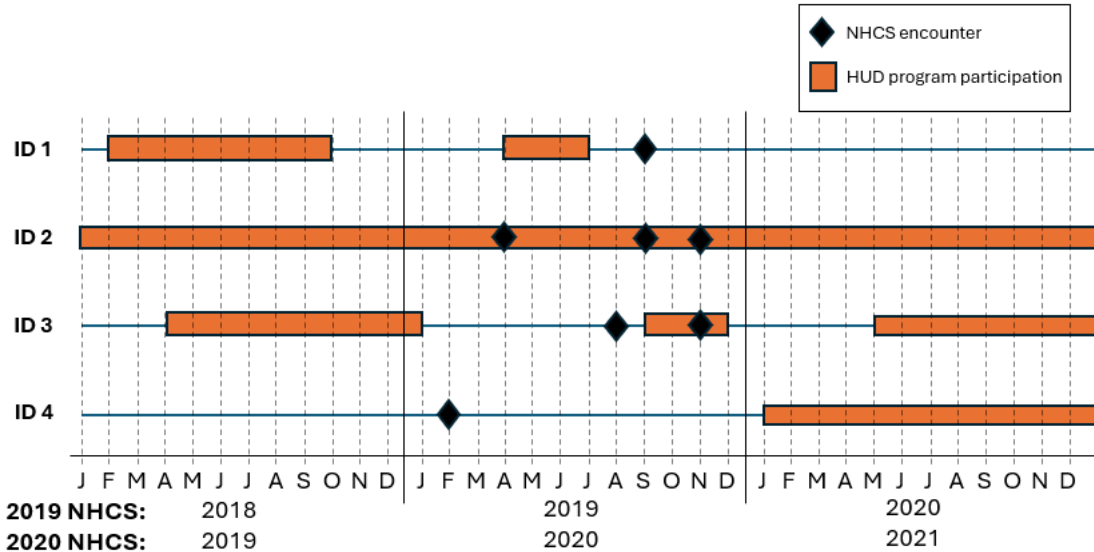
- Patient ID1 did receive HUD assistance during the NHCS data collection period but their HUD program participation period was not concurrent with the month (September) in which Patient ID1 had a hospital encounter.
- Patient ID2 was receiving HUD assistance for all months in calendar year 2019 or 2020 and had three hospital encounters (April, September, and November) during this period.
- Patient ID3 was receiving HUD assistance during the month of their November hospital encounter but was not receiving HUD assistance during their August hospital encounter.
- Patient ID4 had one hospital encounter (February) but was not receiving HUD assistance at the time of this encounter or anytime during the NHCS data collection period.

To determine which HUD program the patient was participating in at the time of their 2019 or 2020 hospital encounter, the researcher would utilize the monthly indicator variables for 2019 or 2020 found in each of the specific HUD program participation files (HCV, MF, and PH). If a monthly indicator variable is equal to 1, this indicates that the patient was participating in that specific HUD program at the time of their hospital encounter. Note that it is possible for a patient to be participating in more than one HUD program in any given month.

4.2.2.3 Temporal Alignment of HUD Assistance with NHCS Patient Data

The linked data files include monthly HUD program participation variables for three calendar years. Therefore, HUD monthly program participation status may be available for NHCS patients during the calendar year prior to, the year of, or the year after the NHCS data collection period. Expanding on [Figure 1](#), which shows four hypothetical patient encounter and HUD program participation scenarios for the year of data collection, [Figure 2](#) provides examples of how the HUD program participation periods may align with NHCS patient data for the entire HUD linkage period.

Figure 2. Temporal Alignment of 2019 and 2020 NHCS Patient Visit Data Linked to year of, year before and year after HUD Program Participation Data.



Notes: HUD is the U.S. Department of Housing and Urban Development
Sources: Hypothetical example based on NHCS patient data linked to HUD administrative data.

The examples shown in [Figure 2](#) are as follows,

- Patient ID1 was receiving HUD assistance from February through September during the calendar year prior to the data collection period, as well as during the data collection period (April - June); but was not receiving HUD assistance during any month of the calendar year following the data collection period. The one hospital encounter for Patient ID1 was in September of the data collection year and not concurrent with HUD assistance.
- Patient ID2 was receiving HUD assistance throughout the entire NHCS - HUD linkage period (January of year before data collection - December of year after data collection) including at the time of each their hospital encounters.
- Patient ID3 was intermittently receiving HUD assistance with periods of HUD participation prior to (April - December), during (September - November), and after (May - December) the data collection period. The hospital encounter in April of the data collection period was not concurrent with HUD assistance; the encounter in November was concurrent with HUD assistance.
- Patient ID4 was not receiving HUD assistance the year before data collection or at the time of their hospital encounter in February but was receiving HUD assistance for all months of the calendar year following the data collection period.

5 Access to Data Files

5.1 Access to the Restricted-Use Linked NHCS – HUD Administrative Data Files

To ensure confidentiality, NCHS provides safeguards including the removal of all personal identifiers from analytic linked files. Additionally, the linked data files are only accessible through the NCHS Research Data Center (RDC) network for approved research projects. Researchers who wish to access the restricted-use 2019 and 2020 NHCS survey files and the linked NHCS-HUD administrative data files must submit a research proposal application to the NCHS RDC. The RDC staff will review all submitted proposals to determine if the proposed project is feasible and to identify any potential disclosure risks. More information regarding the NCHS RDC network and the RDC proposal application process is available from: <https://www.cdc.gov/rdc/> (accessed July 2025).

5.2 Merging NHCS Survey Data to the Linked NHCS-HUD Administrative Data Files

The linkage between the 2019 and 2020 NHCS data and HUD program data was conducted at a patient level using patient level identifiers. The shared variable, PATIENT_ID, will be used by the RDC to merge the linked 2019 or 2020 NHCS-HUD data files with the restricted-use 2019 or 2020 NHCS data. Analysts should request all variables of interest from the NHCS restricted-use data files and the linked NHCS-HUD data files in their RDC proposal.

5.3 Additional Related Data Sources-Linked NHCS-NDI Mortality Files

In addition to the linked NHCS-HUD Administrative Data files data, researchers may also request variables from the linked 2019 NHCS–2019-2020 NDI data file and the linked 2020 NHCS–2020-2021 NDI data file if mortality is an outcome of interest (NCHS Data Linkage -Restricted-Use Linked NHCS-NDI Data: <https://www.cdc.gov/nchs/linked-data/nhcs/restricted-ndi.html>, accessed July 2025). The linked mortality file includes Patient ID, date of birth, date of death, and cause of death information for linked decedents. To integrate the linked NHCS linked mortality files with the linked NHCS- HUD data files, joins are made on the common identification number, PATIENT_ID.

Appendix I: Detailed Description of Linkage Methodology

1 NHCS and HUD Linkage Submission Files

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

- SSN (validated)⁹
- DOB (month, day, and year)
- Sex
- 5-Digit ZIP code and state of residence
- First, middle initial, and last name

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

- Date values: when invalid or outside of expected range
- Sex values: when multiple sex values are recorded for the same person
- Name values: multiple edits are applied:
 - Removal of special characters such as [“-,<>/?, etc.]
 - Removal of descriptive words such as twin, brother, daughter, etc.
 - Nulling of baby names—name parts that contain specific keywords such as baby, infant, girl or boy are set to null
 - Names listed as Jane/John Doe
 - Removal of titles such as Mister, Miss, etc.
 - Removal of suffixes such as Junior, II, etc.
 - Removal of special text such as first name listed as “Void”

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

- Sex was missing. Two alternate records (one with male sex and the other with female) were created (note that this would result in having generated records run through both male and female specific linkage passes, and resulting duplicated links would be subsequently resolved).
- SSN with less than nine digits. A single alternate record was created where leading zeros were added to SSN values of length 7 or 8 to make a 9-digit SSN. Note, no alternate record was created if an invalid SSN would be created by adding 0’s.
- Improbable date of birth. Age at time of survey was computed by subtracting the year of the survey and the year of birth. Records with age greater than 114 had a single alternate record created,

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

- If month and day were suspected of being imputed (ex. Jan 1st or June 15th), entire DOB was changed to missing¹⁰
- Otherwise, only year was changed to missing
- State of residence outside of U.S. and not in rest of world (RW) list. Alternate record was created with ZIP and state codes changed to missing
- ZIP code represents a different state. Using the ZIPSTATE() SAS function, state was imputed using the non-missing ZIP code. If the imputed state was different from the recorded state of residence, an alternate record using imputed state was created
- Multiple name parts and common nicknames (see below)

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

Table 2. Example of Alternate Record Generation using Name Fields

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

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

Submission files, which combined the cleaned and validated PII fields, were created separately for NHCS patient records and for HUD enrollment records. During this process, multiple submission records were created for each patient/enrollee to show all combinations of the recorded values for these fields. That is, if a patient/enrollee had two states-of-residence recorded and three dates-of-birth recorded and each of the remaining fields had only one variant, then a total of six submission records would have been created for the patient/enrollee (see [Table 3](#) for example). Submission records that did not meet the eligibility requirements (see [Section 3.1](#) Linkage Eligibility Determination) were removed from the submission file.

¹⁰ Note, the date values are often recorded when the actual value is unknown.

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

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

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

Additional post processing steps were taken after the initial NHCS and HUD linkage submission files were created. First, records from both the NHCS and HUD submission files were separated according to the sex value (male or female). As mentioned in [section 3.2](#), the probabilistic linkage method assumes independence between the PII variables used to score the potential links. Records in the submission files were separated by sex to avoid violating this assumption, especially when first and/or last name and sex would be used as blocking and/or scoring variables. Additionally, the HUD submission file is limited to records with an effective date between January 1, 2018 and December 31, 2020 for the 2019 NHCS linkage and January 1, 2019 and December 31, 2021 for the 2020 NHCS linkage. This step was taken to reduce the computational burden of linking records that will ultimately be rejected because they occur outside of the linkage period.

2 Deterministic Linkage Using Unique Identifiers

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

3 Probabilistic Linkage

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

3.1 Blocking

Blocking is a key step in the probabilistic record linkage process. It identifies a smaller set of potential candidate pairs, eliminating the need to compare every single pair in the full comparison space (i.e., the Cartesian product). According to Christen, blocking or indexing, “splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key).”¹¹ Intuitively developed rules can be used to define the blocking criteria, however, for this linkage, variable values in the data being linked were used to inform the development of a set of blocking passes that efficiently join the datasets together (i.e., multiple, overlapping blocking passes are run, each using a different blocking key). By using these data to create an efficient blocking scheme (or set of blocking passes), a high percentage of true positive links were retained while the number of false positive links were significantly reduced. A supervised machine learning algorithm used the ‘truth source’ (see [Appendix I section 2](#)) as the validation dataset and the NHCS and HUD submission records as training data. For more detailed information on the supervised machine learning algorithm used, please refer to “Learning Blocking Schemes for Record Linkage” and “Using supervised machine learning to identify efficient blocking schemes for record linkage”.^{12 13}

The machine learning algorithm produced 14 blocking passes to be used in the blocking scheme. [Table 4](#) provides the PII variables that were assigned to each of the blocking passes and the PII variables that were used to score the potential links in each of the blocking passes. Note, the variables listed in the scoring key are all PII variables not used as a blocking variable. Further, if only the ZIP code of residence was used as a blocking variable, then state of residence was excluded from the list of scoring variables as it is implied to agree on all records.

¹¹ Christen, Peter. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Berlin Heidelberg: Springer-Verlag, 2012. <http://www.springer.com/us/book/9783642311635> (accessed June 2025).

¹² Michelson, Matthew, and Craig A. Knoblock. “Learning Blocking Schemes for Record Linkage.” In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, 440–445. AAAI’06. Boston, Massachusetts: AAAI Press, 2006. <https://pdfs.semanticscholar.org/18ee/d721845dd876c769c1fd2d967c04f3a6eaa.pdf> (accessed June 2025)).

¹³ Campbell, S. R., Resnick, D. M., Cox, C. S., & Mirel, L. B. (2021). Using supervised machine learning to identify efficient blocking schemes for record linkage. *Statistical Journal of the IAOS*, 37(2), 673–680. <https://doi.org/10.3233/SJI-200779> (accessed June 2025).

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

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

3.2 Score Pairs

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

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

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

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

Except for first and last name, agreement status was set to 1 if the NHCS and HUD values for a particular PII variable agreed exactly, 0 if they disagreed, and missing (i.e., '.') if either value was missing on the paired records. The agreement status assignment for first and last name is explained further in [Section 3.2.2](#) of this appendix.

3.2.1 M and U Probabilities

The M-probability is the probability that the identifiers on a pair of records agree, given that records represent the same person (i.e., the records are a match). M-probabilities were estimated separately within each individual blocking pass and were calculated for each of the identifiers used for scoring ([Table 4](#)). Within the blocking pass, pairs with agreeing SSN were used to calculate the M-probabilities, as these are assumed to represent the same individual. SSN agreement was defined as having 8 or more digits being the same for pairs with a full 9-digit SSN or the last 4-digits being the same for pairs with only a 4-digit SSN (ex. XXXXX9999). Further, to account for the alternate submission records generated during the creation of the submission files, the “best” agreement was taken for each of the scoring variables among the blocked records for each NHCS patient ID and HUD ID (see [Tables 5](#) and [6](#) for example of alternate record summarization). [Table 5](#) is an example of how the agreement flags for each of the scoring variables in Blocking pass 10 are created. A value of 1 means the information in the variable is exactly matching, while a 0 means they are not. [Table 6](#) then represents how the multiple submission records in [Table 5](#) are summarized into one record for each NHCS patient and HUD administrative ID. If any of the identifiers agree across multiple records, they are flagged as agree (i.e., set to 1). The summarized records in [Table 6](#) are then used to estimate the M-probabilities for each of the specific scoring variables.

Table 5. Example of Agreement Flags Using Blocking Pass 10

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

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

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

Table 6. Example Showing Summarization of Blocked Record Pairs for M-Probability Estimation, based on Table 5 example

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

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

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

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

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

The U-probability is the probability that the two values for an identifier from paired records agreed given that they were NOT a match. Similar to the M-probabilities, U-probabilities were calculated only for the PII variables not included in the blocking keys and with the exception of first and last names, were computed within the blocking pass. The U-probabilities were computed using records where non-missing SSNs were not in agreement (defined as having less than 5 matching digits when records had a full 9-digit SSN and less than 4 matching digits for records with a 4-digit SSN). In order to avoid skewing U-probabilities in blocking passes that contained a high percentage of deterministic matches, assumed matches (i.e., records where SSN was not in agreement and had majority of the non-missing PII among scoring variables

in agreement) were excluded prior to calculating the U-probabilities. For example, when computing the U-probability for day of birth in blocking pass 12, record pairs that did not agree on SSN that had a majority (i.e., greater than 50%) of the PII among first name, middle initial, and month of birth in agreement were excluded from the assumed non-matches. Even though SSN did not agree, these records were assumed to be probable links given that a majority of the PII between the NHCS and HUD submission records agreed.

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

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

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

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

3.2.2 M and U Probabilities for First and Last Names

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

Because of the large number of unique name values, it was impractical to compute name specific U-

probabilities for each blocking pass (i.e., there would not be enough records available for it to be done accurately). Instead, U-probabilities were estimated using pairs generated by the Cartesian product of all records in the NHCS linkage submission file and 100,000 randomly selected names from a simple random sample of 10% of records with non-missing name information from the HUD submission file. See [Table 7](#) for the number of sampled HUD submission records.

Table 7. Count of Records from a 10% Simple Random Sample of 1999-2022 HUD Submission Records used to Estimate U-Probabilities for First and Last Names by Sex

Sex	Count of Sampled Records by Name	
	First Name	Last Name
Female	3,412,344	3,427,331
Male	2,231,647	2,243,904

Complete name tallies (separately, for first and last names) were then produced for the NHCS linkage submission file. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. For each Jaro-Winkler level, the number of names in agreement of the 100,000 randomly selected HUD file names were then tallied.^{14 15 16}

3.2.3 Adjustment of U-Probabilities for Alternate Submission Records

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

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

¹⁴ 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.

¹⁵ 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.

¹⁶ Resnick, D., Mirel, L., Roemer, M., & Campbell, S. (2020). Adjusting Record Linkage Match Weights to Partial Levels of String Agreement. *Everyone Counts: Data for the Public Good.* Joint Statistical Meetings (JSM). <https://ww2.amstat.org/meetings/jsm/2020/onlineprogram/AbstractDetails.cfm?abstractid=312203> (accessed July 2025).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

NOTES: Data have been fabricated for the purposes of this example. JW is the Jaro-Winkler string comparator function.

¹ The collapsed U-probability includes only the U-probabilities after similar names have been collapsed into a single name.

² The adjusted U-probability is computed by summing each of the collapsed 85% JW U-probabilities within each patient ID.

³ The Maximum U-probability is the max U-probability value between the original and adjusted 85% U-probabilities.

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

3.2.4 Calculate Agreement and Non-Agreement Weights

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

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

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

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

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

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

3.2.5 Calculate Pair Weight Scores

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

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

First name and last name weights were assigned using Jaro-Winkler similarity scores described in [Section 3.2.2](#). These scores ranged from 0 to 1, with 0 representing no similarity and 1 representing exact agreement. The weighting algorithm assigned all similarity scores 0.85 and below 0.85 a disagreement weight. The algorithm assigned all similarity scores above 0.85 an agreement weight associated with the 0.85 level. If there was an agreement at the 0.85 level, the algorithm assessed the pair at the 0.90 level given that it agreed at the 0.85 level. If the names disagreed at this level, the algorithm assigned them a

disagreement weight (specific to the 0.90 level given agreement at the 0.85 level). If the names agreed, the algorithm assigned them an additional agreement weight (specific to the 0.90 level). This process continued two more times: for the 0.95 and 1.00 thresholds.

3.3 Probability Modeling

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

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

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

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

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

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

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

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

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

Continuing with the example from Step 1...

if for Pair 1 of blocking pass B, the pair-weight is 8.4, then $Odds_{1,B} = 2^{(8.4 + -5.61)} \approx 6.9$

if for Pair 2 of blocking pass B, the pair-weight is -2.5, then $Odds_{2,B} = 2^{(-2.5 + -5.61)} \approx 0.0036$

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

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

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

Continuing with the example...

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

For Pair 2 in blocking pass $P_{EM,P,B}(Match) = \left(\frac{0.0036}{0.0036 + 1} \right) \approx 0.0036$

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

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

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

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

$$N_{matches,B} = 0.87 + .0036 + P_{EM,3,B} + \dots + P_{EM,N_{pairs,B},B}$$

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

3.4 Adjustment for SSN Agreement

Up to this point, every pair generated through the probabilistic routine was assigned a value that estimates its probability of being a match. However, this estimate did not take SSN agreement into account. This was conducted as a separate step because for the other comparison variables, M- and U- probabilities were estimated based on probable matches or non-matches that were determined based on SSN agreement, and clearly this was infeasible for SSN itself.¹⁸

To remedy this, before the algorithm adjudicated the matches against the probability cut-off value, one final adjustment was made to the match probabilities (for probabilistic pairs). For pairs that had an SSN on both the NHCS and HUD submission record, the estimated probability was adjusted based on the last four

digits of the SSN.

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

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

When the last four digits of SSN did not agree:

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

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

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

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

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

4.1 Estimating Linkage Error to Determine Probability Cut-off Value

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

- Type I Error: Among pairs that are linked, what percentage of them were not true matches.
- Type II Error: Among true matches, how many were not linked.

Because all records were included in the probabilistic linkage (i.e., even deterministic links), SSN agreement status (defined as seven or more matching digits for nine-digit SSN's and for SSN's that had only the last four digits, all four digits must match) was used to measure Type I error. Type I error for probabilistic links was measured as the total number of probabilistic links with non-agreeing SSN divided by the total number of probabilistic links with a valid SSN available on both the NHCS and HUD submission record. Also, deterministically established links were considered to have 0% Type I error rates. While it was believed that the error for these links was quite small and near 0, it is expected that some error does exist even with the deterministically established links and so the estimate was likely biased low. For example, if 40% of links were derived from the probabilistic method, this would reduce the estimated Type I error by the proportion of probabilistically determined linkages among all linkages. To further illustrate, if the Type I error rate for probabilistic links was estimated as 1.2%, then the estimated Type I error rate for the combined linkage process would be $(0.40 * 0.012) = 0.0048$ or 0.48%.

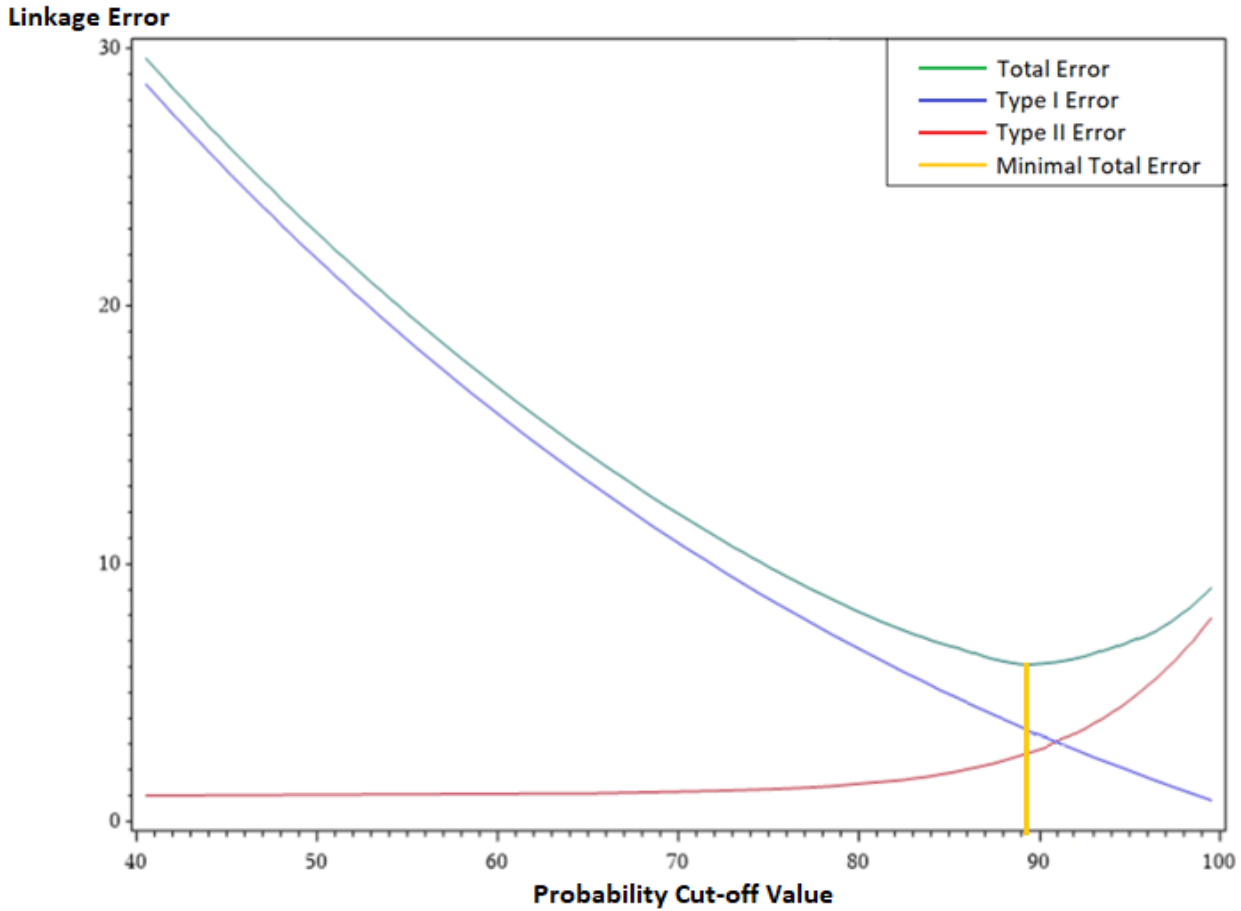
¹⁸ The M and U probabilities in the formulas refer specifically to the M and U of the last four digits of the SSN.

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

4.2 Set Probability Cut-off Value

One goal of record linkage is to have the lowest errors possible. However, as more pairs are accepted, pairs that are less certain to be matches but accepted as links increase the Type I error and decrease Type II error. And as less pairs are accepted, pairs that are more certain to be matches but not accepted as links decrease the Type I error and increase Type II error. The optimal trade-off between Type I error and Type II error is not known, but it can be assumed to be optimal when the sum of Type I and Type II error is at a minimum. For this reason, Type I and Type II error are estimated at various probability cut-off values and the one that showed the lowest estimate of total error is selected (see [Figure 3](#)). The optimal probability cut-off value was set to 0.8925 for both the 2019 NHCS and 2020 NHCS linkage to HUD data.

Figure 3. Illustrating linkage error by probability cut-off value
 (Illustrative schematic not based on actual values)



4.3 Select Links Using Probability Cut-off Value

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

4.4 Resolving NHCS Patient IDs that Linked to Multiple HUD Records

Due to the nature of the 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 2019 NHCS data, 44.8% of patients were linked to more than one HUD enrollment record with the same HUD ID. This number was slightly lower for the 2020 NHCS at 41.8%. In situations where a NHCS patient ID linked to more than one HUD enrollment record with different HUD IDs, and the PROBVALID score calculated for each unique linked enrollment record exceeded the 0.8925 probability cut-off value, all HUD ID matches were assumed to represent the same individual. In the 2019 NHCS data, about 1.5% of linked patients were linked to more than one HUD ID. This number was slightly lower for the 2020 NHCS at 1.4%. For more information on how to use PROBVALID values to reduce potential Type 1 errors see [Section 4.2.1.1](#).

4.5 Computed Error Rates of Selected Links

Final error rates were computed for selected links (described in [Section 4.3](#)). [Table 11](#) provides the total number of selected links, the number of total links identified through deterministic and probabilistic methods, and the Type I and Type II error rates for the 2019 and 2020 NHCS HUD linkages. Because the links were selected using the SSN adjusted probability (described in [Section 4.1](#)), the overall Type I error rate was computed using the estimated match probabilities rather than using SSN agreement. For the probabilistic links, the estimated match probabilities represented the probability that the 2019 or 2020 NHCS patient record was a match to the HUD administrative record. In other words, if a link had an estimated probability of 0.98, then it was understood that there was a 98% chance this was a match. To estimate the Type I error rate for the probabilistic links, the chance that a link is not a match was summed (i.e., $\sum 1 - Prob_{valid_SSN_Adj}$) and then divided by the total number of probabilistic records. The method to measure the overall Type II error remained unchanged (see [Section 4.1](#)).

Table 11. Algorithm Results for Total Selected Links

	Probability Cut-off Value	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
2019 NHCS	0.8925	283,668	72,339	211,329	0.14%	1.52%
2020 NHCS	0.8925	319,067	86,087	232,980	0.14%	1.47%