National Center for Health Statistics
Data Linkage

The Linkage of the 2014 and 2016 National Hospital Care Survey to U.S. Department of Housing and Urban Development Administrative Data: Matching Methodology and Analytic Considerations

Data Release Date: August 31, 2021

Document Version Date: August 31, 2021

Division of Analysis and Epidemiology

National Center for Health Statistics

Centers for Disease Control and Prevention

datalinkage@cdc.gov

Suggested Citation: National Center for Health Statistics. Division of Analysis and Epidemiology. *The Linkage of the National Hospital Care Survey to U.S. Department of Housing and Urban Development Administrative Data: Matching Methodology and Analytic Considerations*, August 2021. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data-linkage/index.htm

Contents

List of Acronyms

CCD, Continuity of Care Document

CMS, Centers for Medicare & Medicaid Services

DOB, date of birth

ED, emergency department

EHR, electronic health record

EM, expectation-maximization

ERB, Ethics Review Board

HCV, Housing Choice Voucher program

IP, inpatient

MBSF, Master Beneficiary Summary File

MF, Multi-family housing programs

MTW, Moving to Work 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.htm (accessed August 6, 2021). The NHCS collects electronic health records or health care claims data from participating hospitals drawn from a national sample frame of 581 hospitals. Participating hospitals are requested to send all patient ambulatory care and inpatient (IP) encounters occurring within the data collection calendar year. The NHCS includes detailed information about each participating hospital's patients' characteristics, conditions, and treatment. Even though NHCS is an establishment survey (i.e., hospitals are the sampling unit) it collects patient personally identifiable information (PII), which enable data linkages.

Through its Data Linkage Program, NCHS has been able to expand the analytic utility of the data collected from NHCS by augmenting it with housing assistance program data collected by the U.S. Department of Housing and Urban Development (HUD). **This report will describe the linkage of the 2014 NHCS to 2013-2015 HUD administrative data and the 2016 NHCS to 2015-2017 HUD administrative data.** Although the 2014 and 2016 NHCS data are not nationally representative due to low survey response rates, linking NHCS with HUD administrative data creates a new data resource that can support research studies focused on a wide range of patient health outcomes and the role of housing assistance programs as a social determinant of health.

This report includes a brief overview of the 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 in part through contract #HHSD2002016F92236B by NORC at the University of Chicago with funding from the Department of Health and Human Services' Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF).

2 Background on Linked Files

2.1 National Hospital Care Survey

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

From participating hospitals, NHCS collects data on all IP and ambulatory care visits occurring during the calendar year. In 2014, hospitals were required to provide data from claims records, but to reduce the burden of reporting on participating hospitals, for the 2016 data collection hospitals were given the option of providing their data in the form of electronic health records (EHRs) or as claims records. Thus, participating hospitals provided data in the form of Uniform Bill (UB)-04 administrative claim records or EHR data, where the EHR data are provided in the form of Continuity of Care Documents (CCDs) or custom extracts.

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

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

2.2.1 HUD Public and Assisted Housing Programs

The U.S. Department of Housing and Urban Development (HUD) is the primary federal agency responsible for overseeing domestic housing programs and policies. While HUD is responsible for administrating various housing and community development programs, the linkage with the 2014 and 2016 NHCS focuses on HUD's three largest housing assistance programs: Housing Choice Vouchers (HCV), Public Housing (PH), and Multifamily programs (MF). Persons and households participating in these program types are "HUD-assisted."

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

The HUD Housing Choice Voucher (HCV) program is the federal government's largest housing assistance program, allowing low-income families, elderly persons, and persons with disabilities to choose and lease safe and affordable housing. In the HCV program, housing assistance is

tenant-based, meaning that participants find their own housing in the private market. Participants are free to choose any housing that meets program requirements and are not limited to units located in subsidized housing projects. In the NHCS-HUD linked data, the HCV program also includes the Homeownership Voucher, Project-Based Voucher, Section 8 Moderate Rehabilitation, and Section 8 Rental Certificate programs. Among 2014 and 2016 NHCS patients that linked to HUD administrative data, just over 50% were participating in an HCV program.

The multifamily (MF) program category in the linked NCHS–HUD data encompasses a number of separate, distinct HUD programs, including: Project-Based Section 8 (or PBS8) Voucher Assistance in Multifamily Housing (the largest MF program), Section 221(d)(3) Below Market Interest Rate, Section 236 Multifamily Housing, Rental Assistance, Section 202 Supportive Housing for the Elderly Program, Section 202/162—Project Assistance Contract, Section 811 Supportive Housing for Persons with Disabilities, and Rent Supplement. Because each of the remaining MF programs lacked sufficient sample size on an individual basis in the linked file, they were combined into a single MF program category. In all MF programs, subsidies are paid directly to private property owners who provide a certain percentage of their housing units at affordable rates for low-income persons who qualify. MF program assistance is tied to the property, unlike tenant-based rental assistance programs (e.g., HCVs), and tenants cannot take their rental housing assistance subsidy elsewhere. Approximately 25% of the 2014 and 2016 NHCS patients that linked to HUD were participating in a MF program.

The 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. HUD also provides technical assistance to help PHAs plan, develop, and manage PH developments. Approximately 25% of the 2014 and 2016 NHCS patients that linked to HUD were participating in a PH 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 PIC data extract created for the NHCS-HUD data linkage was based on HUD's PIC point-in-time quarterly files, which capture a household's most recent transaction with HUD during the prior 18 months (with the exception of Moving to Work (MTW) demonstration program participants, where 36 months is used as the threshold). A transaction refers to any activity for which a HUD form was completed (e.g., new admission to a HUD program, annual recertification, end of participation, etc.). These files are released four times a year.

TRACS is a system developed to collect and maintain certified tenant data from owners and management agents of MF housing programs. 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. These data capture transactions occurring within the 18 months immediately prior to the date of extract. Transactions with the same SSN, effective date, and transaction code were considered duplicates and removed.

To determine program overlap, HUD transactions collected from PIC and TRACS were used to create participation episodes for the final linked NHCS-HUD administrative 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 August 6, 2021).

3 Linkage Methodology

3.1 Linkage Eligibility Determination

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

Linkage was attempted only for NHCS patient records that had at least two of the following three identifiers present: valid SSN[2], valid date of birth (month, day, and year)[3] or valid name (first, middle initial, and last)[4]. For example, if the PII on the NHCS patient record had no SSN, a full name, and only the year of birth, the record would be considered ineligible for linkage, as only one of the criteria (i.e., that for name) was met.

The variable ELIGSTAT, included on the linked NHCS-HUD match file, provides the linkage eligibility status (which indicates whether the linkage eligibility criteria had been met) for each NHCS patient record. ELIGSTAT values include 0 (ineligible) or 1 (eligible). Table 1 presents the total number of 2014 and 2016 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 percentage of total sample and eligible for linkage who were linked to HUD administrative program 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.

[1] The NCHS ERB, also known as an Institutional Review Board or IRB, is an appointed ethics review committee that is established to protect the rights and welfare of human research subjects.

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

[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, and last) are non-missing.

3.2 Overview of Linkage

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

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-Sunter method.[5] Following this, a selection process was implemented with the goal of selecting pairs believed to match (i.e., representing the same individual between the data sources).

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

For each NHCS patient-level record that was linked, HUD extracted information from the PICS and TRACS systems and sent them to NCHS through a secure data transfer system. Table 7 highlights the linkage results by deterministic and probabilistic links.

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

Table 1. Linked NHCS – HUD Administrative Records: Sample Sizes and Percent Linked, by Age and Sex

	Sample Size			Percent Linked	
	Total Sample	Eligible for Linkage[3]	Linked to HUD Administrative Data[4]	Total Sample[5]	Eligible Sample[6]
2014 NHCS					
Age[1]					
0-17	1,063,289	961,790	100,939	9.5	10.5
18-44	1,155,989	1,050,841	86,242	7.5	8.2
45-61	630,731	574,740	40,557	6.4	7.1
62 and over	707,187	640,714	41,448	5.9	6.5
Total	3,557,196	3,228,085	269,186	7.6	8.3
Sex[2]					
Male	1,577,255	1,434,577	92,104	5.8	6.4
Female	1,978,165	1,791,197	176,941	8.9	9.9
Total	3,555,420	3,225,774	269,045	7.6	8.3
2016 NHCS					
Age[1]					
0-17	1,293,458	1,205,473	122,502	9.5	10.2
18-44	1,477,611	1,386,926	112,874	7.6	8.1
45-61	796,022	748,333	54,309	6.8	7.3
62 and over	888,601	836,014	56,013	6.3	6.7
Total	4,455,692	4,176,746	345,698	7.8	8.3
Sex[2]					
Male	2,597,453	1,851,201	116,174	4.5	6.3
Female	3,157,461	2,278,263	225,141	7.1	9.9
Total	5,754,914	4,129,464	341,315	5.9	8.3

NOTES: Data are presented at patient level.

[1] Age is as of final IP or ED encounter (date of last known contact). Age could not be determined for 1,090 patients in the 2014 NHCS and for 1,367,473 patients in the 2016 NHCS due to missing data. Age is calculated by subtracting patient date of birth (DOB) from the final encounter date. When more than one DOB was present, the minimum of the non-missing DOB was selected.

[2] Sex could not be determined for 2,866 patients in the 2014 NHCS and for 68,251 in the 2016 NHCS due to missing data.

[3] Eligibility for linkage is based upon having sufficient PII in at least two of three data element groups: SSN, name, and date of birth. 330,104 patient records in the 2014 NHCS and 1,642,060 in the 2016 NHCS were missing all PII and were also considered ineligible for linkage.

[4] This group includes linkage-eligible patients who linked to HUD enrollment database at any time during the linkage interval (2014 NHCS: 2013 – 2015 HUD, 2016 NHCS: 2015 – 2017 HUD).

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

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

4 Analytic Considerations

This section summarizes some key analytic 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 linked NHCS-HUD data files are encouraged to read "A Primer on HUD Programs and Associated Administrative Data" for additional information on HUD program and corresponding administrative data, including important analytic considerations.[6]

4.1 Analytic Considerations for Linked NHCS Data

4.1.1 NHCS Hospital Eligibility and Sampling

Eligible hospitals for NHCS are non-institutional, non-federal hospitals with six or more staffed inpatient beds, and there are 6,622 hospitals which met these criteria as of 2013 to form the NHCS 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 2013, to provide estimates for ED visits with incidents of substance abuse, 81 hospitals with 500 staffed inpatient beds or more were added from the reserve sample. Thus, the hospital sample size for the 2014 and 2016 NHCS data collection (which re-used the 2013 sample) was 581 hospitals. In 2014, 95 of the 581 sampled hospitals provided data. Of the 95, 93 were eligible for linkage (note: this number excludes hospitals that provided less than 50 patient encounter records or patient records covering less than 6 months of the analysis period). Of those 93 participating hospitals, 93 hospitals sent IP data and 82 hospitals sent ED data. In 2016, 158 out of the 581 sampled hospitals provided data and of the 158, 142 hospitals were eligible for linkage (note: this number excludes hospitals that did not provide patient PII or provided less than 50 patient encounter records or did not provide patient records covering at least 6 months of the analysis period). Of those 142 participating hospitals, 140 hospitals sent IP data and 121 hospitals sent ED data.

4.1.2 NHCS Sampling Weights Are Currently Not Available

Currently, there are no sampling weights available for the 2014 or 2016 NHCS data. This section will be updated if sampling weights are made available in the future. Because the hospital level sampling conducted for the NHCS was not conducted on an equal probability basis, unweighted estimates will be biased to be more similar to those from hospitals selected with higher sampling probability. Similarly, there will be bias towards types of hospitals responding at higher rates. These biases will be more of a concern if estimates vary strongly by factors correlated with sampling and response rates.

One way to mitigate these biases in the absence of survey weights is to calculate estimates in the framework of regression modeling that controls for hospital characteristics. This would be done by including hospital characteristics (region, ownership type, and size) as well as patient

[6] https://www.cdc.gov/nchs/data/datalinkage/primer-on-hud-programs.pdf (accessed August 6, 2021)

characteristics (age and sex) among the predictor variables in the model definition. Statistical testing can then be conducted on parameter estimates associated with these characteristics.

4.1.3 NHCS Patient Identification Number

Each patient in the NHCS is assigned a unique identification number, PATIENT_ID. PATIENT_ID does not contain any identifiable information about the patient and is intended to be unique for each individual receiving IP, ED, or OPD services at a participating hospital. However, the de-duplication of patient records required to generate this ID depends on sometimes incomplete or erroneous data, 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.[7]

4.2 Analytic Considerations for Linked HUD Data Files

4.2.1 Description of NCHS-HUD Linked 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 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.

Data linkages include some uncertainty over which pairs represent true matches. An estimated probability of match validity (PROBVALID) was computed for each candidate pair and compared against a probabilistic cut-off value to determine which pairs were links (an inferred match). For additional discussion on how PROBVALID was estimated, see Appendix I, Sections 3.3 and 3.4. NCHS used a probabilistic cut-off value which minimized the total estimated counts of Type I error (false positive links – identified as 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 HUD Match file, NCHS used a probabilistic cut-off value of 0.9225 to determine final match status. Candidate pairs with a PROBVALID that exceeded the probabilistic cut-off (i.e., PROBVALID>0.9225) were deemed a link. For additional discussion on cut-off determination and record selection please see Appendix I, Section 4. For some analyses, it may be desirable to reduce the Type I error. In order to do this, researchers should increase the probability cut-off value (to a value closer to 1.0). Of note, the PROBVALID cannot be decreased from 0.9225. Researchers wishing to access PROBVALID in order to change the link acceptance cut-off value, should request this variable in their RDC proposal.

[7] For more information on Patient_ID generation, see Technical Notes on page 14: https://www.cdc.gov/nchs/data/nhsr/nhsr097.pdf (accessed August 6, 2021)

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 a least one HUD housing assistance program administrative transaction record during the linkage period.

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 at the time of the hospital encounter, as well as the calendar year prior to or after the patient encounter. Patients in the 2014 NHCS were linked to HUD program participation transaction records between 2013 and 2015 and patients in the 2016 NHCS were linked to HUD program participation transaction records between 2015 and 2017.

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 created for each NHCS, including a summary program participation file (Linked_HUD_Pgrm_NHCS20XX_Any) 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 NHCS patient ID and 36 monthly HUD participation indicators (one for each month during the linked data time span). For example, the monthly HUD enrollment indicators in the linked 2014 NHCS-HUD HCV program specific participation file are HCV_JAN2013 through HCV_DEC2015. The monthly indicators are created from program participation episodes that were derived using the transaction dates from the transaction file of matches extracted by HUD. For each month between the start and end date (including the start and end dates) of the participation episode 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. Of note, some HUD program participation transaction periods began prior to or ended after the NHCS-HUD linkage period.

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 value of one in HUD_MATCH_STATUS indicates that an NHCS patient ID was linked with a HUD record at least once during the linkage period. In order to determine which of the specific HUD programs the patient was participating in, researchers should use the program specific participation files (one for each of the three main HUD programs). Each program-specific file contains monthly indicators where a value of one indicates patient program participation for that month.

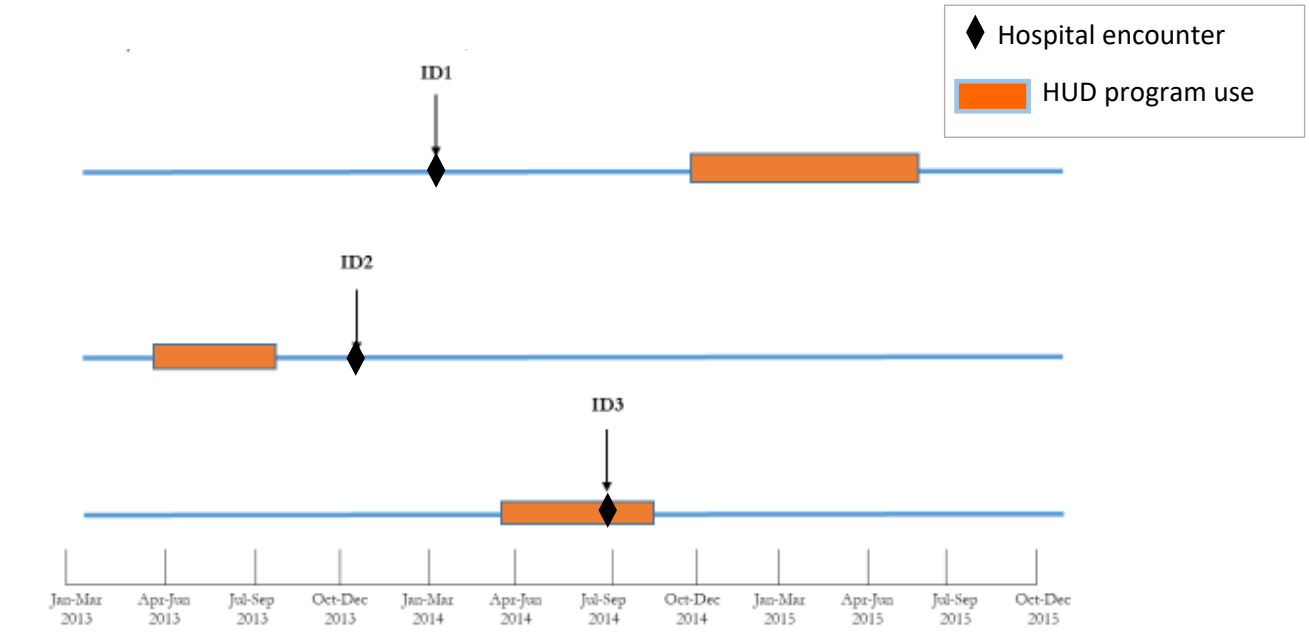
4.2.2.2 Temporal Alignment of HUD Assistance

To identify whether the NHCS patient was enrolled in a HUD program during, before, or after their hospital encounter, researchers can compare the month and year of the NHCS hospital

encounter, typically using the date of admission, with the monthly indicators included on any of the linked HUD program participation data files.

Figure 2 below depicts three possible scenarios related to the temporal alignment of HUD housing assistance and patient encounter data for three hypothetical patients noted as patient ID1, patient ID2, and patient ID3. In each timeline, the diamond represents the quarter of the patient encounter and the time period in orange represents the month(s) during which the patient received HUD assistance. Patient ID1 received HUD assistance after their encounter. Patient ID2 received HUD assistance before their encounter. Patient ID3 concurrently received HUD assistance at the time of their encounter.

Figure 2. Temporal alignment of NHCS data linked to HUD administrative data files.



NOTES: HUD is the U.S. Department of Housing and Urban Development
SOURCES: NCHS, 2014 NHCS linked to 2013-2015 HUD administrative data.

For example, as noted in Figure 2, a 2014 NHCS patient with a hospital encounter admission of February 18, 2014 (patient ID1) would be identified as enrolled in a HUD program after their hospital encounter with monthly indicator variables HUD_STATUS_OCT2014 through HUD_STATUS_SEP15 in the summary program participation file (Linked_HUD_Pgrm_NHCS2014_Any) all equal to 1. Patient ID2 had an enrollment in HUD before their hospital encounter and Patient ID3 was enrolled at the time of their hospital encounter. In order to determine which HUD program the patient was participating in at the time of their February 18, 2014 hospital encounter, the researcher would utilize one of the monthly indicator variables (i.e., HCV_FEB2014, PH_FEB2014, or MF_FEB2014) in each of the specific HUD program participation files. If the monthly indicator variable is equal to 1, this indicates that the patient was participating in that specific HUD program at the time of their

encounter. Note that it is possible for a patient to be participating in more than one HUD program in any given month. For example, if a researcher identifies that the HCV_FEB2014 and MF_FEB2014 program specific participation indicators are both equal to 1, this indicates that the patient was participating in both the HCV and MF HUD programs at the time of their February 18, 2014 hospital encounter.

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 August 6, 2021).

5 Access to Data Files

5.0 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 made available in secure facilities for approved research projects. Researchers who wish to access the linked NHCS-HUD administrative data files must submit a research proposal to the NCHS Research Data Center (RDC) to obtain permission to access the restricted use files. All researchers must submit a research proposal to determine if their projects are feasible and to gain access to these restricted data files. The proposal provides a framework which allows RDC staff to identify potential disclosure risks. More information regarding RDC and instructions for submitting an RDC proposal are available from: https://www.cdc.gov/rdc/ (accessed August 6, 2021).

5.1 Merging NHCS Analytic Files to the Linked NHCS-HUD Administrative Data Files

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

The NHCS analytic files include analytically-pertinent hospital-level details (such as bed size and geographic region) and episode-level details (patient demographics, diagnoses, procedures, admission and discharge dates). To perform NHCS patient encounter-level analysis, the linked NHCS - HUD Administrative Data files must be used in conjunction with 2014 and 2016 NHCS analytic files.[8] The shared variable, PATIENT_ID, allows analysts to merge NHCS patient records for the same patients within or across hospital settings (IP or ED) and to merge information from the NHCS - HUD Administrative Data files.

5.2 Additional Related Data Sources – Linked NHCS-NDI Mortality Files

In addition to the NHCS-HUD Administrative Data files data, researchers may also request variables from the 2014 NHCS–2014/2015 NDI linked data file and/or the 2016 NHCS–2016/2017 NDI linked data file if mortality is an outcome of interest (NCHS Data Linkage - Restricted-Use Linked NHCS-NDI Data, accessed August 6, 2021). The linked mortality file includes Patient ID, date of birth, date of death, and cause of death information for linked decedents. To integrate the linked NHCS - NDI linked data files into the linked NHCS- HUD Administrative data files, joins are made on the common identification number, PATIENT_ID.

[8] Find more information about the NHCS analytic files: https://www.cdc.gov/rdc/b1datatype/dt1224h.htm (accessed August 6, 2021)

5.3 Additional Related Data Sources – Linked NHCS-CMS Medicare Files

Researchers interested in analyzing information on HUD housing-assistance and health care utilization for persons also enrolled in Medicare may request variables from the 2014 NHCS – 2014/2015 Center for Medicare & Medicaid Services (CMS) Medicare Master Beneficiary Summary File (MBSF) linked data files and/or the 2016 NHCS – 2016/2017 Medicare Enrollment/Summary, Claims/Encounters, and Assessment Data linked data files (NCHS Data Linkage Restricted-Use Linked NHCS-CMS Medicare Data).

The linked 2014 and 2016 NHCS-CMS Medicare MBSF files include information on Medicare program entitlement and enrollment, summarized annual health care utilization and cost data, and chronic condition flags indicating the presence of certain health conditions for linked Medicare beneficiaries. Additionally, the 2016 NHCS-CMS Medicare linked data files include health care claims and encounters, prescription drug events, and patient assessment data for linked Medicare beneficiaries. To integrate the linked NHCS – CMS Medicare linked data files into the linked NHCS- HUD Administrative 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

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

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

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

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

Similar to the cleaning process, a more elaborate routine was used to generate alternate records involving the name fields. Additional records were generated for patients with multiple name parts, common nicknames, and for common Hispanic and Asian names. NCHS created a common nickname lookup file which was used to generate a second record replacing the nickname with the formal name. Table 2 below provides two examples of how multiple part name information was used to generate alternate records, using hypothetical data. For patient A, the first name was used to generate multiple records, and for patient B, the last name was used.

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

Table 2. Example of Alternate Record Generation using Name Fields

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

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

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

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

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 format SSN. The algorithm performed two passes on the data, first checking for full 9-digit SSN agreement and then for records where the last 4-digits of the SSN agreed. After records had been matched using SSN, the algorithm validated the deterministic links by comparing first name, middle initial, last name, month of birth, day of birth, year of birth, ZIP code of residence, and state of residence. If the ratio of agreeing identifiers to non-missing identifiers was greater than 50% (1st pass using SSN-9) or greater than 2/3 (2nd pass using last 4 of SSN), the linked pair was retained as a deterministic match. Of note, NHCS patients were excluded from the second pass (i.e., using the last 4-digits of SSN) if they were deterministically linked in the first pass. The collection of records resulting from the deterministic match is referred to as the 'truth source.'

3 Probabilistic Linkage

The second step in the linkage process was to perform the probabilistic linkage. To infer which pairs are links, the linkage algorithm first identified potential links and then evaluated their

probable validity (i.e., that they represent the same individual). The following sections describe these steps in detail. The weighting procedure of this linkage process closely followed the Fellegi-Sunter paradigm, the foundational methodology used for record linkage. Based on Fellegi-Sunter, each pair was assigned an estimated probability representing the likelihood that it is a match – using pair weights computed (according to formula) for each identifier in the pair – before selecting the most probable match between two records.

3.1 Blocking

Blocking is a key step in the probabilistic record linkage process. It identifies a smaller set of potential candidate pairs, eliminating the need to compare every single pair in the full comparison space (i.e., the Cartesian product). According to data linkage expert Peter Christen, blocking or indexing, "splits each database into smaller blocks according to some blocking criteria (generally known as a blocking key)."[10] Intuitively developed rules can be used to define the blocking criteria, however, for this linkage, the data being linked were used to inform the development of a set of blocking passes that efficiently join the datasets together (i.e., multiple, overlapping blocking passes are run, each using a different blocking key). By using these data to create an efficient block scheme (or set of blocking passes), a high percentage of true positive links were retained while the number of false positive links were significantly reduced. A supervised machine learning algorithm used the 'truth source' as the validation dataset and a sample of the NHCS and HUD submission records as training data. For more detailed information on the supervised machine learning algorithm used please refer to "Learning Blocking Schemes for Record Linkage."[11,12]

The machine learning algorithm learned 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 be in agreement on all records.

Table 4. Blocking and scoring scheme used to identify and score potential links

Key Number	Blocking Key	Scoring Key
1	Last name, month of birth, day of birth, year of birth	First name, middle initial, state of residence, ZIP code of residence, sex
2	Month of birth, day of birth, year of birth, state of residence, sex	First name, middle initial, last name, ZIP code of residence

[10] 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 August 6, 2021).

[11] Michelson, Matthew, and Craig A. Knoblock. "Learning Blocking Schemes for Record Linkage." In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, 440–445. AAAI'06. Boston, Massachusetts: AAAI Press, 2006. https://pdfs.semanticscholar.org/18ee/d721845dd876c769c1fd2d967c04f3a6eeaa.pdf (accessed August 6, 2021).

[12] 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 August 6, 2021).

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

3.2 Score Pairs

Next, each pair was scored using an approach based on the Fellegi-Sunter paradigm. The Fellegi-Sunter paradigm specifies the functional relationship between agreement probabilities and agreement/non-agreement weights for each identifier used in the linkage process. The scores – pair weights – calculated in this step were used in a probability model (explained in Section 2.3), which allowed the linkage algorithm to select final links to include in the linked file. The scoring process followed the following order:

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

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

- First Name or First Initial (when applicable)
- Middle Initial
- Last Name or Last Initial (when applicable)

- Year of Birth
- Month of Birth
- Day of Birth
- Sex
- State of Residence
- ZIP Code (conditional on state agreement)

3.2.1 Calculate M- and U- Probabilities

The **M-probability** – the probability that the identifiers using the records in question agree, given that records represent the same person – were estimated separately within each individual blocking pass. M-probabilities were calculated for each of the identifiers not used in the blocking key (Table 2). Within the blocking pass, pairs with agreeing SSN (defined as 8 or more digits being the same) were used to calculate the M-probabilities, as these are assumed to represent the same individual. Further, to account for the alternate submission records generated during the creation of the submission files, the "best" agreement was taken for each of the scoring variables among the blocked record for each patient ID and HUD administrative ID (see Tables 5 and 6 for example of record summarization). For example, among qualifying pairs in blocking pass 2, 99.4% agree on day of birth and 94.5% agreed on state of residence. These percentages represented estimates of the M-probabilities for these identifiers.

Table 5. Example of Agreement Flags for Blocked Records

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

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

Table 6. Example Showing Summarization of Blocked Records for M-Probability Estimation

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

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

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

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

The **U-probability** – the probability that the two values for an identifier from paired records agreed given that they were NOT a match. Similar to the M-probabilities, U-probabilities were only calculated for the PII variables not included in the blocking keys and with the exception of first and last names, were computed within the blocking pass. The U-probabilities were computed using records where non-missing SSN were not in agreement (defined as having less than 5 matching digits). In order to avoid skewing U-probabilities in blocking passes that contained a high percentage of deterministic matches, assumed matches (i.e., records where SSN was not in agreement that had majority of the non-missing PII among scoring variables were in agreement) were excluded prior to calculating the U-probabilities. For example, when computing the U-probability for day of birth in blocking pass 12, records that did not agree on SSN that had majority of the PII among first name, middle initial, and month of birth were excluded from the assumed non-matches. These records were assumed to be probable matches given that a majority of the PII between the survey and administrative records were in agreement.

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

3.2.2 M and U Probabilities for First and Last Names

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

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

random sample of 3% (1,474,484 records for first name and 1,481,581 records for last name) of records with non-missing name information of the HUD submission file.

Complete name tallies (separately, for first and last names) were then produced for the NHCS submission file. For each level of name on the file, 100,000 names were randomly selected from the HUD submission file 3% sample to compare to it. Comparisons were made based on the Jaro-Winkler distance metric at four different levels: 1.00 (Exact Agreement), 0.95, 0.90, and 0.85. The number of names in agreeance of the 100,000 randomly selected HUD file names that agreed at that level for each name were then tallied.[13,1415]

3.2.3 Calculate Agreement and Non-Agreement Weights

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

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

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

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

3.2.4 Calculate Pair Weight Scores

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

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

First name and last name weights were assigned using Jaro-Winkler similarity scores described in Section 3.2.2. These scores ranged from 0 to 1, with 0 representing no similarity and 1 representing exact agreement. The weighting algorithm assigned all scores below 0.85 a

[13] 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.
[14] 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.
[15] 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 August 6, 2021).

disagreement weight. The algorithm assigned all scores above 0.85 an agreement weight associated with the 0.85 level. If there was an agreement at the 0.85 level, the algorithm assessed the pair at the 0.90 level *given* that it agreed at the 0.85 level. If the names disagreed at this level, the algorithm assigned them a disagreement weight (specific to the 0.90 level given agreement at the 0.85 level). If the names agreed, the algorithm assigned them an additional agreement weight (specific to the 0.90 level). This process continued two more times: for the 0.95 and 1.00 thresholds.

3.3 Probability Modeling

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

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

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

1. A pair-weight adjustment was computed (Adj_B) specific to blocking pass, *B*, by taking the log base 2 of the estimated number of matches (within blocking pass *B*) divided by the estimated number of non-matches in the blocking pass. For convenience, the estimated number of matches, $\widehat{N_{matches,B}}$, 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, out of the number of pairs, $N_{Pairs,B}$ = 1,000,000, then

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

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

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

   Continuing with the example from Step 1…

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

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

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

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

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

   Continuing with the example…

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

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

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

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

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

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

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

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

3.4 Adjustment for SSN Agreement

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

non matches that were determined based on SSN agreement and clearly this was infeasible for SSN itself.[16]

To remedy this, before the algorithm adjudicated the matches against the probability threshold, one final adjustment was made to the match probabilities (for probabilistic pairs). For pairs that had an SSN on both the NHCS and HUD administrative record, the estimated probability was adjusted based on the last four digits of the SSN.[17]

When the last four digits of SSN[18] agreed (i.e., are exactly the same):

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

When the last four digits of SSN did not agree:

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

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

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

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

4.1 Estimating Linkage Error to Determine Probability Cutoff

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

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

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

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

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

- Type II Error: Among true matches, how many were not linked

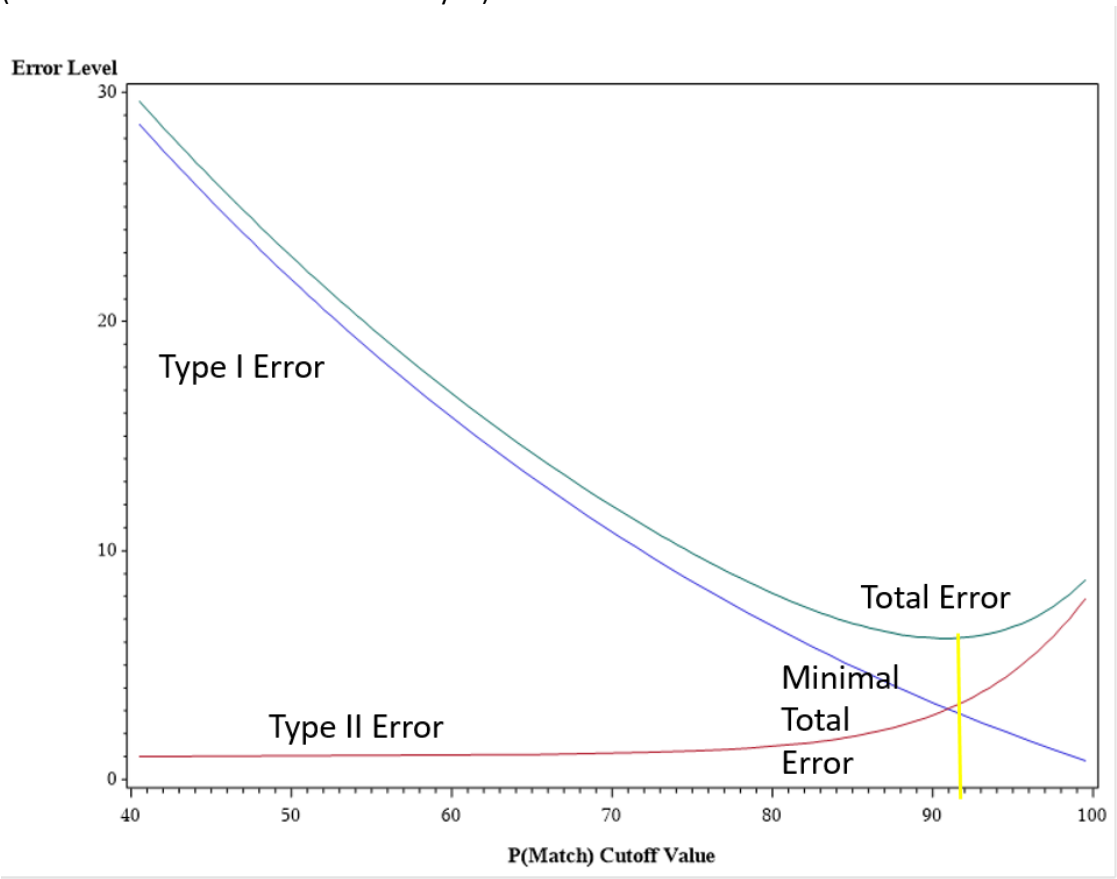
Because all records were included in the probabilistic linkage (i.e., even deterministic links), SSN agreement status (defined as 7 or more matching digits) was used to measure Type I error. Type I error for probabilistic links was measured as the total number of probabilistic links with non-agreeing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

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

4.2 Set Probability Cutoff

One goal of record linkage is to have the lowest errors possible. However, as more pairs were accepted, pairs that were less certain to be matches as links increase the Type I error and decrease Type II error (see Figure 3). And as less pairs were accepted, pairs that were more certain to be matches as links decrease the Type I error and increase Type II error. The optimal trade-off is between Type I error and Type II error was not known, and likely this depends on the type of analysis to be conducted with the linked data, but it is assumed that it is not far from optimality when the sum of Type I and Type II error is at a minimum. For this reason, Type I and Type II error are estimated at various probability cut points and the one that showed the lowest estimate of total error was selected. For this linkage, the probability cutoff was set to 0.9225.

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



4.3 Select Links Using Probability Threshold

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

4.4 Resolving NHCS Patient ID that Linked to Multiple HUD Enrollment Records

Due to the nature of administrative program data, it is possible that PII information may vary, due to PII changes over time or recording errors, among HUD enrollment records that represent the same person. In the 2014 NHCS, 23.3% of patients were linked to more than one HUD enrollment record with the same HUD ID and 20.5% of 2016 NHCS patient records were similarly linked. In situations where a NHCS patient ID linked to more than one HUD enrollment record with different HUD IDs, and the PROVALID score calculated for each unique linked enrollment record exceeded the 0.9225 cutoff value, all HUD ID matches were assumed to represent the same individual. In the 2014 and 2016 NHCS, about 2% of linked patients were linked to more than one HUD ID. 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 7 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 2014 and 2016 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 NHCS 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 (1 - $Probvalid_{SSN_{Adj}}$) and then divided by the total number of probabilistic records. The method to measure the overall Type II error remained unchanged (see Section 4.1).

Table 7. Algorithm Results for Total Selected Links by Year of NHCS

	Cutoff	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
2014 NHCS	0.9225	336,354	98,034	238,320	0.1%	2.2%
2016 NHCS	0.9225	422,920	124,328	298,592	0.1%	1.9%

Table 8 provides the total selected links, number of probabilistic and deterministic links, and the estimated Type I and II error rates for the selected links, by survey year and by record type source for the 2016 NHCS. Note: All hospitals participating in the 2014 NHCS provided data in the form of UB-04 claims. As shown in Table 8, UB-04 Claims have higher estimated linkage error (both Type I and II) compared to the EHR records. Due to elevated levels of missing data in EHRs compared to the UB-04 claims records, the number of deterministic matches made by the algorithm for EHR Custom Extract (66.3%) is proportionally higher than UB-04 deterministic matches (24.9%). This resulted in a lower proportion of EHRs having HUD administrative data extracted based on the probabilistic linkage. Additionally, CCD data were delivered without SSN

information. This resulted in 100% of CCDs having HUD administrative data extracted based on the probabilistic linkage and therefore the Type II linkage error rate was not calculated.

Table 8. Algorithm Results for Total Selected Links by 2016 NHCS Data Source

Data Source	Cutoff	Total Selected Links	Deterministic Matches	Probabilistic Links	Est Incorrect (Type I)	Est Not Found (Type II)
UB-04 Claims	0.9225	318,545	79,434	239,111	0.1%	2.5%
EHR Custom Extract	0.9225	67,753	44,894	22,859	<0.1%	0.6%
CCD	0.9225	36,622	0	36,622	0.2%	*

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