



The Linkage of National Center for Health Statistics Survey Data to United States Renal Data System (USRDS) End-Stage Renal Disease (ESRD) Patient Data

Linkage Methodology and Analytic Considerations

Data Release Date: November 12, 2021
Document Version Date: October 28, 2021

Suggested citation:

National Center for Health Statistics, Division of Analysis and Epidemiology. The Linkage of National Center for Health Statistics Survey Data to United States Renal Data System (USRDS) End-Stage Renal Disease (ESRD) Patient Data - Methodology and Analytic Considerations. November 2021. Hyattsville, Maryland.

Available at the following address: <https://www.cdc.gov/nchs/data-linkage/esrd.htm>

Table of Contents

1 Introduction.....	4
2 Data Sources.....	4
2.1 National Center for Health Statistics (NCHS)	4
2.2 United States Renal Data System (USRDS) End-Stage Renal Disease (ESRD) Information.....	6
3 Linkage of NCHS Surveys with 1974 through 2018 USRDS ESRD Patient Data	8
3.1 Linkage Eligibility	8
3.2 Linkage Methods.....	9
3.3 Data Confidentiality	13
4 Analytic Guidelines and Considerations	13
4.1 General Analytic Guidance.....	13
4.2 Merging Linked NCHS-USRDS ESRD Data with NCHS Survey Data	14
4.3 Sample Weights	14
4.4 Timing of Data.....	15
4.5 Inconsistencies with dates.....	15
4.6 Sample sizes	15
5 Analyses that Have Used the Linked NCHS-USRDS ESRD Data	16
6 References.....	17
Appendix: Constructing Unique Identification Numbers used to Merge Datasets.....	18
1 National Health Interview Survey (NHIS), 1994-2018.....	18
1.1 NHIS, 1994.....	18
1.2 NHIS, 1995-1996	18
1.3 NHIS, 1997-2003	19
1.4 NHIS, 2004.....	19
1.5 NHIS, 2005-2018	19
2 National Health and Nutrition Examination Survey (NHANES), 1999-2018	20
3 Third National Health and Nutrition Examination Survey (NHANES III)	20
4 National Nursing Home Survey (NNHS), 2004	20

List of Abbreviations

CDC, Centers for Disease Control and Prevention
CKD, chronic kidney disease
CMS, Centers for Medicare and Medicaid Services
ERB, Ethics Review Board
ESRD, end-stage renal disease
HICN, health insurance claim number
LSOA II, Second Longitudinal Study of Aging
MEC, mobile examination center
MPP, Medicare primary payer
MSP, Medicare secondary payer
NCHS, National Center for Health Statistics
NHANES, National Health and Nutrition Examination Survey
NHIS, National Health Interview Survey
NHEFS, NHANES I Epidemiologic Follow-up Study
NIDDK, National Institute of Diabetes and Digestive and Kidney Diseases
NNHS, National Nursing Home Survey
PII, personally identifiable information
RDC, Research Data Center
RESNUM, resident record number
SSN, social security number
USRDS, United States Renal Data System

The Linkage of the National Center for Health Statistics Survey Data to United States Renal Data System (USRDS) End-Stage Renal Disease (ESRD) Patient Data – Methodology and Analytical Considerations

1 Introduction

Federally sponsored health surveys are a critical source of public health information in the United States. The National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention is the nation's principal health statistics agency and is responsible for collecting accurate, relevant, and timely health data. NCHS conducts several national surveys and collects vital statistics data to monitor the health of the American people. These national health surveys provide rich information on topics such as health conditions and behaviors, health insurance coverage, access to health care, and socioeconomic status, but longitudinal information on outcomes is often not available. Demand is increasing to incorporate information from additional sources to enhance the availability and quality of information on exposures and outcomes and to provide longitudinal information on healthcare utilization or health outcomes when available.

NCHS' Data Linkage Program is designed to maximize the value of NCHS population-based surveys by augmenting survey information with health-related information from [death certificates](#), health care utilization data from [Medicare](#) and [Medicaid](#), and housing data from [federal housing assistance programs](#). Linked data files enable researchers to examine the factors that influence disability, chronic disease, health care utilization, morbidity, and mortality.

Under a designated agent agreement between NCHS and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), data from several NCHS population-based surveys have been linked to end-stage renal disease (ESRD) (i.e., permanent kidney failure requiring dialysis or a kidney transplant) patient data obtained from the United States Renal Data System (USRDS). Even though the ESRD population remains at less than 1% of the total Medicare population, it has accounted for about 7% of Medicare fee-for-service spending in recent years (1). The linked NCHS-USRDS ESRD files allow for the examination of factors related to chronic kidney disease (CKD) and ESRD among NCHS survey participants. For some survey participants, ESRD data may be available during the year the NCHS survey was conducted. Data may also be available in the years prior to or after the survey. This linkage includes ESRD data through 2018, updating and replacing previous data linkages available at <https://www.cdc.gov/nchs/data-linkage/esrd.htm>. Given the low prevalence of ESRD among children and the potential disclosure risk, the linked files only include adults from the NCHS surveys (2).

2 Data Sources

2.1 National Center for Health Statistics (NCHS)

Adults from following NCHS surveys were included in the linkage to USRDS ESRD patient data:

- 1994-2018 National Health Interview Survey (NHIS)
- 1999-2018 National Health and Nutrition Examination Survey (NHANES)
- Third National Health and Nutrition Examination Survey (NHANES III)
- The Second Longitudinal Study of Aging (LSOA II)
- NHANES I Epidemiologic Follow-up Study (NHEFS)
- 2004 National Nursing Home Survey (NNHS)

A brief description of the NCHS surveys follows:

NHIS is a nationally representative, cross-sectional household interview survey that serves as an important source of information on the health of the civilian, non-institutionalized population of the United States. It is a multistage sample survey with primary sampling units of counties or adjacent counties, secondary sampling units of clusters of houses, tertiary sampling units of households, and finally, persons within households. It has been conducted continuously since 1957 and the content of the survey is periodically updated. Beginning in 1997, NHIS implemented a questionnaire redesign to obtain more detailed health information for selected individuals within a sampled household. Within each household, families are identified, and a family respondent completes a brief structured interview on family demographics and broad health measures. Starting in 1997, from each family in NHIS, one adult aged 18 or over (sample adult) and—if present—one child (sample child) is randomly selected, and information on each is collected with the Sample Adult Core and the Sample Child Core interviews. The content of these two interviews differs on some items, but both collect basic information on health status, health care services, and health behaviors. For the Sample Adult Core interview, the selected individual responds for himself or herself (i.e., no proxy response is allowed, except when the person is unable to respond due to a physical or mental condition) (3). For the Sample Child Core interview, an adult who is knowledgeable about the sample child is the respondent. NHIS has been used as the sampling frame for other NCHS surveys focusing on specialized populations, including LSOA II. For detailed information on the NHIS contents and methods, refer to the NHIS website: <https://www.cdc.gov/nchs/nhis/index.htm> (accessed October 28, 2021).

LSOA II was a prospective study of a nationally representative sample of civilian, noninstitutionalized persons 70 years of age or older at the time of their 1994 NHIS interview, which served as the baseline for the study. The LSOA II study design included two follow-up telephone interviews, conducted in 1997-1998 and 1999-2000. The LSOA II provides information on changes in disability and functioning, individual health risks and behaviors in the elderly, and use of medical care and services employed for assisted community living. For detailed information on the LSOA II contents and methods, refer to the LSOA II website: <https://www.cdc.gov/nchs/lsao/lsao2.htm> (accessed October 28, 2021).

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

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

NHEFS was a national longitudinal study conducted in collaboration with the National Institutes of Health, National Institute on Aging, and other agencies of the Public Health Service. The NHEFS cohort included all persons 25–74 years of age who completed a medical examination as part of NHANES I in 1971–75. The NHEFS study design included four follow-up interviews, conducted in 1982–84, 1986, 1987 and 1992, to investigate the relationships between clinical, nutritional, and behavioral factors assessed at baseline and subsequent morbidity, mortality, and institutionalization.

For detailed information about the Continuous NHANES, NHANES III and NHEFS contents and methods, refer to the NHANES website: <https://www.cdc.gov/nchs/nhanes/index.htm> (accessed October 28, 2021).

NNHS provides information on nursing homes from two perspectives—that of the provider of services and that of the recipient of care. Data for the surveys were obtained through personal interviews with facility administrators and designated staff who used administrative records to answer questions about the facilities, staff, services, and programs. Medical records were also used to answer questions about the residents. NNHS was first conducted in 1973–1974 and repeated in 1977, 1985, 1995, 1997, 1999, and most recently in 2004. Only the 2004 survey was included in this linkage. For more information on the NNHS content and methods, refer to the NNHS website, <http://www.cdc.gov/nchs/nnhs.htm> (accessed October 27, 2020).

2.2 United States Renal Data System (USRDS) End-Stage Renal Disease (ESRD) Information

The USRDS is a national data system funded by NIDDK, designed to collect, analyze, and distribute information about ESRD in the United States. The linked NCHS survey and USRDS ESRD data files can be used by researchers interested in conducting analyses related to patients with ESRD. Much of the ESRD information that follows is based on the USRDS website, <https://www.usrds.org/for-researchers/> (accessed October 28, 2021).

ESRD data are divided into five files:

- Patient Profile
- Medical Evidence Report (form 2728)
- Detailed Treatment History
- Condensed Treatment History
- Payer History

The **Patient Profile (PATIENTS) File** contains basic demographic and ESRD-related data including primary and contributing causes of ESRD diagnosis, date of first dialysis, and information on kidney transplants. It also contains death information if a Death Notification form 2746 was completed. Date of first ESRD service (FIRST_SE) is a variable that may be of particular interest to researchers. FIRST_SE is non-missing for all patients that linked, and it can be used to estimate both the incidence and prevalence of ESRD among NCHS survey participants at different points in time. The Patient Profile File has one record for each linked NCHS survey participant. The data dictionary for the file can be found at: <https://www.cdc.gov/nchs/data-linkage/esrd-restricted.htm> (accessed October 28, 2021).

The **Medical Evidence Form (MEDEVID) File** contains data from 1987, 1995, 2005, and 2015 versions of the Centers for Medicare and Medicaid Services' (CMS) Medical Evidence Report (CMS-2728) Form. This is the source of data regarding primary cause of kidney failure and start date of chronic dialysis. The CMS 2728 form is used to register patients at the onset of ESRD and must be submitted by dialysis or transplant providers within 45 days of treatment initiation. It establishes Medicare eligibility for individuals who previously were not Medicare beneficiaries, reclassifies previously eligible Medicare beneficiaries as ESRD patients, and provides demographic and diagnostic information for all new ESRD patients regardless of Medicare entitlement. CMS, USRDS, and kidney research communities rely on the form to ascertain patient demographics, primary diagnosis, comorbidities, and biochemical test results at the time of ESRD onset. Because the form has changed over time, the response categories found in the USRDS ESRD data may vary based on year of collection. The notes section of the Medical Evidence Report Form data dictionary details how information is collected on different versions of this form. The file may contain multiple records for each linked NCHS survey participant. The data dictionary for the Medical Evidence Form File can be found at: <https://www.cdc.gov/nchs/data-linkage/esrd-restricted.htm> (accessed October 28, 2021).

The **Detailed Treatment History (RXHIST) File** records the sequence of treatment modalities for each ESRD patient. Each record in the file indicates a period of therapy with a given modality. Consequently, any change in provider or detailed modality results in a new record. Variables include start and end date of modality period, detailed treatment modality description, and a flag that indicates whether the end date coincided with the patient's death. The file may contain multiple records for each linked NCHS survey participant. The data dictionary for the Detailed Treatment History File can be found at: <https://www.cdc.gov/nchs/data-linkage/esrd-restricted.htm> (accessed October 28, 2021).

The **Condensed Treatment History (RXHIST60) File** is similar to the Detailed Treatment History File, but it does not include detailed treatment modality or the USRDS assigned facility identification. The file may contain multiple records for each successfully linked NCHS survey participant entered into USRDS. The data dictionary for the Condensed Treatment History File can be found at: <https://www.cdc.gov/nchs/data-linkage/esrd-restricted.htm> (accessed 10/28/2021).

The **ESRD Payer History (PAYHIST) File** contains a continuous sequential history of payers for each patient in the ESRD database, beginning with the first ESRD service date. Each patient's first service data in the Payer History File is the same date reported in the Treatment History File. Data from the Medicare Enrollment Database and dialysis claims information are used to categorize payer status as Medicare primary payer (MPP), Medicare secondary payer (MSP), or non-Medicare. The claims database contains data only for MPP and MSP patients, so claims-based analyses, such as analyses of cost and hospitalization, will not include non-Medicare patients. Non-Medicare patients, therefore, must be identified and excluded when determining numbers of patients or patient years at risk for analyses of cost per patient or hospitalization rates. Also, as it is impossible to determine the complete hospitalization history or complete cost of care for ESRD patients with MSP coverage, such analyses should also exclude patients during the periods when they have non-Medicare coverage. The Payer History File can be used to make these exclusions. The ESRD Payer History file may contain multiple records for each linked NCHS survey participant. The data dictionary for the Payer History File can be found at: <https://www.cdc.gov/nchs/data-linkage/esrd-restricted.htm> (accessed October 28, 2021).

3 Linkage of NCHS Surveys with 1974 through 2018 USRDS ESRD Patient Data

3.1 Linkage Eligibility

The linkage of NCHS survey participants data to USRDS ESRD patient data was conducted under a designated agent agreement between NCHS and NIDDK. Approval for the linkage was provided by NCHS' Research Ethics Review Board (ERB) and the linkage was performed only for eligible NCHS survey participants. The NCHS ERB, also known as an Institutional Review Board or IRB, is an administrative body of scientists and non-scientists that is established to protect the rights and welfare of human research subjects. For the NCHS-USRDS ESRD linkage, only adult NCHS survey participants who provided consent as well as their full or partial social security number (SSN) or full or partial Medicare Health Insurance Claim Number (HICN) and other necessary personally identifiable information (PII), were considered linkage-eligible. Linkage eligibility refers to the potential ability to link data from an NCHS survey participant to administrative data. Criteria for NCHS-USRDS ESRD linkage eligibility vary by survey and year due to the variability of questions across NCHS surveys, changes to PII collection procedures by the surveys over the linkage time period, and changes in which survey respondents are asked linkage related questions.

For the 1994-2006 NHIS, LSOA II, 1999-2008 NHANES, NHANES III, NHEFS, and 2004 NNHS, a refusal by the survey participant to provide an SSN or Medicare HICN was considered an implicit refusal for data linkage. However, over time NCHS noticed an increase in the refusal rate for providing SSN and HICN, particularly for the NHIS, which reduced the number of survey participants eligible for linkage (4). To address declining linkage eligibility rates, NCHS introduced new procedures for obtaining linkage consent from survey participants. Research was also conducted to assess the accuracy of matching data from NHIS to the National Death

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

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

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

3.2 Linkage Methods

A primarily deterministic (rules-based) approach was followed to link the NCHS survey data with USRDS ESRD patient records. The linkage was conducted in two phases. In Phase I, USRDS attempted to match records from the NCHS submission file to the records in the USRDS database to determine the preliminary linkage status for each linkage-eligible survey participant. In Phase II, NCHS determined the final linkage status for each linked pair (NCHS record and USRDS record) based on additional identifiers that matched between the two data sources. The linked NCHS-USRDS ESRD files only contain data that were determined to be final links for linkage-eligible survey participants based on the procedures described in this section. The overall linkage process is described below and illustrated in Figure 1.

Phase I: Determining Preliminary Linkage Status

NCHS prepared and securely transferred to USRDS a submission file containing the following data elements for linkage-eligible survey participants:

- SSN9 (nine-digit SSN) or SSN4 (last four digits of SSN)*
- HICN (if available)
- Date of birth (month, day, year)
- Name (first and last)
- Sex

*note for linkage eligible participants, if HICN was provided SSN9 or SSN4 was extracted from the Medicare HICN only if the survey participant was identified as the primary claimant for Medicare benefits.

The file used for linkage did not contain the NCHS survey public-use identification number, nor did it contain any information that could identify the original survey source. The public-use identification number was replaced with an encrypted linkage identification number used by NCHS staff for data linkage projects.

To determine the preliminary linkage status for each survey participant, USRDS staff attempted to match the records from the NCHS submission file to the USRDS database in four passes using the following identifying information:

Pass 1: SSN9

Pass 2: HICN

Pass 3: SSN4

Pass 4: Date of birth (month, day, year), sex, first name (non-missing) and last name (non-missing)

The results of each pass were output to a separate file. The records on each file included indicators identifying the data elements from the NCHS submission record that matched the USRDS record in addition to the matching variable(s) which defined each pass (e.g., Pass 1: SSN9). The four files were returned to NCHS to determine the final linkage status.

Phase II: Determining Final Linkage Status

NCHS used match indicator variables provided by USRDS to determine the final linkage status for each preliminary link. Data from a recent linkage of NCHS survey data and Medicare administrative data was also incorporated into the adjudication process. Additional information about these linkages can be found at: https://www.cdc.gov/nchs/data-linkage/cms/nchs_medicare14_18_linkage_methodology_and_analytic_considerations.pdf and https://www.cdc.gov/nchs/data-linkage/cms/nchs_medicare_linkage_methodology_and_analytic_considerations.pdf (accessed October 28, 2021). For the preliminary links on the Pass 4 file, ESRD status was confirmed using the ESRD indicator available in the linked NCHS-Medicare data. The rules applied by NCHS for determining final linkage status are listed below.

Pass 1 Match

Records that satisfied one of the following rules were considered a confirmed link in pass 1:

- A. SSN9 match + matched on birth month, birth year, and sex
- B. SSN9 match + at least three of the following data elements matched: birth month, birth year, sex, first name (non-missing), last name (non-missing)

Pass 2 Match

Records that satisfied one of the following rules were considered a confirmed link in pass 2:

- C. HICN match + matched on birth month, birth year, and sex
- D. HICN match + at least three of the following data elements matched: birth month, birth year, sex, first name (non-missing), last name (non-missing)

Pass 3 Match

Records that satisfied one of the following rules were considered a confirmed link in pass 3:

- E. SSN4 match + **ALL** of the following data elements matched: birth month, birth year, sex, first name (non-missing), and last name (non-missing)
- F. SSN4 match + **ALL** of the following data elements matched: birth month, birth year, sex, first name (non-missing), and day of birth
- G. SSN4 match + **ALL** of the following data elements matched: birth month, birth year, sex, last name (non-missing), and day of birth

Pass 4 Match

Records that satisfied the following rule were considered a confirmed link in pass 4:

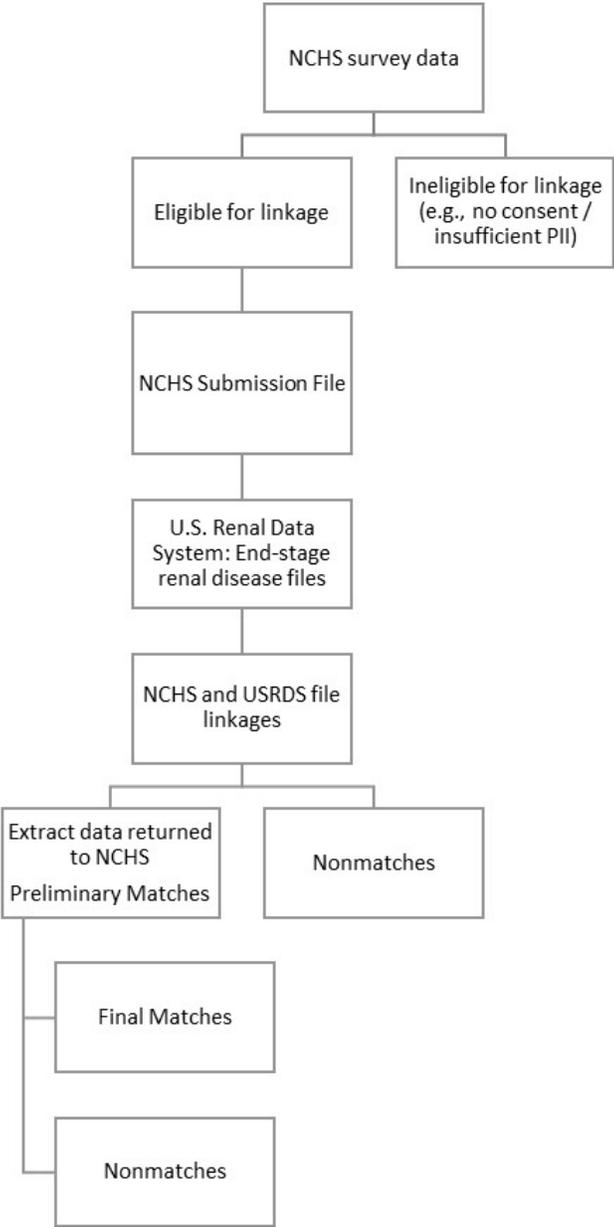
- H. Date of birth, sex, name match + confirmation of ESRD status in the linked NCHS-Medicare data

An additional adjudication step was required to select a single (final) linked USRDS record per survey participant since some survey participants may have had records that linked in multiple passes. In situations where multiple linked records were identified for a survey participant, the final linked record was determined by selecting the first record for each survey participant in the following order of link confidence based on the rules above:

1. Pass 1 match - Rule A
2. Pass 2 match - Rule C
3. Pass 1 match - Rule B
4. Pass 2 match - Rule D
5. Pass 3 match - Any rule (E, F, or G)
6. Pass 4 match - Rule H

If there were ties based on the rules above, the number of matching fields were counted and the record with the highest number of matching fields was selected as the match. (Note: date of birth was weighted by 1/3 for month, day, year.)

Figure 1. Linkage process: linked NCHS – USRDS ESRD data



3.3 Data Confidentiality

NCHS must provide safeguards for the confidentiality of its survey participants. All personal identifiers have been removed from the linked NCHS-USRDS ESRD data files to ensure confidentiality. However, because of the small possibility of reidentification, the linked NCHS-USRDS ESRD data are not available as public-use files.

Researchers who want to analyze the linked NCHS-USRDS ESRD data must submit a research proposal to the NCHS Research Data Center (RDC) to obtain permission to access the restricted use files. Please see the NCHS RDC website for instructions on submitting a proposal: <https://www.cdc.gov/rdc/index.htm> (accessed October 28, 2021).

4 Analytic Guidelines and Considerations

This section describes general considerations and guidelines for analysis using the linked data files identified by NCHS programmers and analysts. Should any new analytic issues be discovered during analysis, notifications can be reported to the NCHS Data Linkage Methodology and Analysis Branch at datalinkage@cdc.gov.

4.1 General Analytic Guidance

To create analytical files for use in the RDC, a researcher provides a file containing variables from the public-use NCHS survey data to the RDC for merging with the NCHS-USRDS ESRD files. Researchers can also request access to restricted variables from NCHS surveys. Each restricted-use variable needs to be specifically requested as part of a researcher's application to the RDC. This allows RDC staff to verify the full list of variables (restricted and public-use) and check for potential disclosure risk.

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

- Geography—Users who require information on geography should request these data in their RDC proposal. Geographic information is also available on the administrative data for linked participants and may differ from that in the survey.
- Linked mortality data for NCHS surveys—Data from eligible survey participants from all NCHS surveys that have been linked to the USRDS ESRD data have also been linked to mortality files. For survey participants that died, the linked mortality files include date and cause of death obtained from death certificate data. These linked data may be useful to researchers and must be specifically requested as part of the researcher's proposal to the RDC.
- NHANES month and year of examination and interview—NHANES data are released in 2-year cycles. The exact year (and month) of a participant's interview and examination is not provided on public-use files. However, researchers will

want to know the time elapsed between a given year (or even month) of the USRDS ESRD data and the NHANES interview or examination. The restricted-use variables that indicate the month and year of NHANES interview or examination must be specifically requested.

- Match status— The match status variable indicates if a survey participant was eligible for linkage and whether they linked to USRDS ESRD data. This variable can be used to adjust sample weights for linkage eligibility.

Researchers are advised to request the following variables, available from the public-use NCHS survey files, for inclusion in analytic files:

- NCHS sample weights and design variables—Sample weights are needed to create nationally representative estimates. Similarly, design variables are required to account for the survey design in analyses. The names of the weights and design variables differ depending on which NCHS survey is being used. These can be identified using the documentation for each NCHS survey. Additional information regarding how to adjust the survey weights for linkage eligibility is provided below.
- Demographic information about survey participants from the NCHS survey is self- or family participant-reported and, thus, may be more accurate than demographic data provided in USRDS ESRD files. Therefore, where possible, NCHS suggests including the demographic data as collected from the survey when analyzing the linked data.

4.2 Merging Linked NCHS-USRDS ESRD Data with NCHS Survey Data

To perform person-level analysis, the restricted-use Linked NCHS-USRDS ESRD Data Analytic Files can be used in conjunction with the NCHS survey data. A unique identification number is available on each file that allows analysts to merge data collected from survey participants with their information from the NCHS-USRDS ESRD Linked Data files. The unique survey identification numbers are survey-specific and may be constructed differently across survey years. Please refer to the appendix for guidance on identifying and constructing (if necessary) the appropriate identification variable for merging survey data and the NCHS-USRDS ESRD Linked Data Files.

4.3 Sample Weights

The sample weights provided in NCHS population health survey data files adjust for oversampling of specific subgroups and differential non-response and are post-stratified to annual population totals for specific population domains to provide nationally representative estimates. The properties of these weights for linked data files with incomplete linkage, due to ineligibility for linkage and nonmatches, are unknown. In addition, methods for using the survey weights for some longitudinal analyses require further research. Because this is an important

and complex methodological topic, ongoing work at NCHS and elsewhere is examining the use of survey weights for linked data.

Until specific recommendations are available, preliminary guidance is to analyze linked data files using linkage-adjusted sample weights. The sample weights available on NCHS population health survey data files can be adjusted for linkage eligibility. This adjustment involves calculating estimates of numbers of people in categories of interest and determining cell sizes by related categories. This process accounts for the potential bias due to linkage eligibility. More detailed information on adjusting sample weights for linkage eligibility using SUDAAN is available in Appendix III from: https://www.cdc.gov/nchs/data/series/sr_01/sr01_058.pdf Researchers should be aware that all NCHS surveys linked to the USRDS ESRD administrative data have complex survey designs. Therefore, considerations of statistical power should account for the survey design, in addition to the unweighted number of observations available for a particular project. Given the small number of participants that linked to the USRDS ESRD data, analysts should use caution when studying certain population subgroups. For additional guidance see the NCHS Data Presentation Standards for Proportions: https://www.cdc.gov/nchs/data/series/sr_02/sr02_175.pdf (accessed October 28, 2021).

4.4 Timing of Data

The NCHS survey data and USRDS ESRD data included in this linkage were collected at different times. Specifically, the NCHS survey data were collected between 1971 and 2018 and ESRD data were collected between 1974 and 2018. For more information see pages 4-6 and page 8 of this report. The linked data files provide the opportunity to examine the administrative data during the year the survey was conducted, in years following the survey, and the years prior to the survey for some NCHS survey participants.

4.5 Inconsistencies with dates

Because USRDS ESRD information comes from administrative data sources, the potential exists for some variables to include values that are outside of the expected range of dates. Data users should anticipate the potential for outliers and include data cleaning in their plans for analyzing the data.

4.6 Sample sizes

The surveys that have been linked along with the sample sizes in the linked USRD ESRD files can be found in the publicly-available sample size tables: <https://www.cdc.gov/nchs/data-linkage/esrd-methods.htm> (accessed October 28, 2021. Because the linkage described in this report replaces previous NCHS-USRDS ESRD linkages, only data from this linkage are available to researchers. Before submitting a proposal to the RDC, researchers are also encouraged to view the publicly-available match rate tables: [NCHS Data Linkage - USRDS End-Stage Renal](#)

[Disease Data - Linkage Methods \(cdc.gov\)](https://www.cdc.gov/disease-data-linkage-methods/) (accessed October 28, 2021) for the surveys they wish to analyze.

5 Analyses that Have Used the Linked NCHS-USRDS ESRD Data

A citation list of scientific articles based on linked NCHS-USRDS ESRD Data Files as of December 22, 2020, can be found below:

1. Crews DC, et al. Race/ethnicity, dietary acid load, and risk of end-stage renal disease among US adults with chronic kidney disease. *Am J Nephrol.* 2018;47(3):174-181.
2. Banerjee T, et al. Food insecurity, CKD, and subsequent ESRD in US adults. *Am J Kidney Dis.* 2017;70(1):38-47.
3. Banerjee T, et al. High dietary acid load predicts ESRD among adults with CKD. *J Am Soc Nephrol.* 2015;26(7):1693-1700.
4. Muzaale AD, et al. Risk of end-stage renal disease following live kidney donation. *JAMA.* 2014;311(6):579-586.
5. Melamed ML, et al. 25-hydroxyvitamin D levels, race, and the progression of kidney disease. *J Am Soc Nephrol.* 2009;20(12):2631-2639.

Technical Reports

1. Golden C, et al. Linkage of NCHS population health surveys to administrative records from Social Security Administration and Centers for Medicare & Medicaid Services. National Center for Health Statistics, 2015. *Vital Health Stat* 1(58).

6 References

1. United States Renal Data System. (2018) Annual Data Report. Chapter 9: Healthcare Expenditures for Persons with ESRD. https://www.usrds.org/media/1734/v2_c09_esrd_costs_18_usrds.pdf
2. Chand DH, Swartz S, Tuchman S, Valentini RP, Somers M. Dialysis in Children and Adolescents: The Pediatric Nephrology Perspective. *American Journal of Kidney Diseases* 69(2):p278-8. 2017.
3. Parsons VL, Moriarty C, Jonas K, Moore TF, David KE, Tompkins L. Design and estimation for the National Health Interview Survey, 2006 – 2015. *National Center for Health Statistics. Vital Health Stat* 2(165). 2014. Miller DM, Gindi R, Parker JD.
4. Trends in record linkage refusal rates: Characteristics of National Health Interview Survey participants who refuse record linkage. Presented at Joint Statistical Meetings 2011. Miami, FL, July 30–August 4.
5. Sayer B, Cox CS. How many digits in a handshake? National Death Index matching with less than nine digits of the social security number in Proceedings of the American Statistical Association Joint Statistical Meetings. 2003.
6. Dahlhamer JM, Cox CS. Respondent consent to link survey data with administrative records: results from a split-ballot field test with the 2007 National Health Interview Survey. Paper presented at the 2007 Federal Committee on Statistical Methodology Research Conference, Arlington, VA, 2007.

Appendix: Constructing Unique Identification Numbers used to Merge Datasets

The data provided on the 1994-2018 NHIS, 1999-2018 NHANES, NHANES III, and the 2004 NNHS linked USRDS ESRD files can be merged with the NCHS restricted and public use survey data files using the unique survey-specific public identification number (PUBLICID/SEQN/RESNUM). Information on how to identify and/or construct the NCHS survey specific PUBLICID, SEQN or RESNUM is provided below.

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

1.1 NHIS, 1994

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

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

SAS example:

`length publicid $14;`

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

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

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

1.2 NHIS, 1995-1996

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

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

SAS example:

`length publicid $14;`

```
PUBLICID = trim(left(YEAR| |HHID| |PNUM));
```

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

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

1.3 NHIS, 1997-2003

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

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

SAS example:

```
length publicid $14;
```

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

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

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

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

1.4 NHIS, 2004

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

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

SAS example:

```
length publicid $14;
```

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

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

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

1.5 NHIS, 2005-2018

Public-use

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

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

SAS example:

`length publicid $14;`

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

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

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

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

<u>Item</u>	<u>Length</u>	<u>Description</u>
SEQN	6	Participant identification number

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

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

<u>Item</u>	<u>Length</u>	<u>Description</u>
SEQN	5	Participant identification number

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

4 National Nursing Home Survey (NNHS), 2004

<u>Item</u>	<u>Length</u>	<u>Description</u>
RESNUM	6	Resident Record (Case) Number

All of the 2004 NNHS public-use data files are linked with the common resident record (case) number (RESNUM). Merging information from the 2004 NNHS Files to the 2004 NNHS-USRDS ESRD linked files

using this variable ensures that the appropriate information for each survey participant is linked correctly.