# Assessing Data Quality in Population-Based Surveys Linked to End-Stage Renal Disease Administrative Data

**Authors: Jonathan Aram[1], Crescent B. Martin[1], Lisa B. Mirel[1]**
[1]Centers for Disease Control and Prevention, National Center for Health Statistics

Proceedings of the 2022 Federal Committee on Statistical Methodology Research and Policy Conference

## Background

Integrating data creates new resources that can be used to answer complex health and policy-related questions that cannot be answered by a single data source alone. The National Center for Health Statistics (NCHS), one of 13 principal federal statistical agencies and the principal US statistical agency for health, undertakes data integration as part of its mission. Through its Data Linkage Program, NCHS links data from NCHS surveys and administrative records to enhance both sources of information. Recently, NCHS updated the linkage of two population-based surveys, the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES), to End-Stage Renal Disease (ESRD) data from the United States Renal Data System (USRDS). The survey data provide information about topics such as access to health care, health insurance, obesity, nutrition, and hypertension, among others; and the ESRD data provide information on the occurrence and treatment of ESRD, and payment for ESRD care. This analysis evaluates how closely prevalence estimates generated using the linked NCHS-USRDS ESRD data align with national prevalence estimates reported in the USRDS Annual Report, with the overarching goal of assessing the accuracy and reliability of the linked data. In addition, the analysis highlights considerations including reduced sample sizes and the potential to introduce new sources of bias in estimates when not all survey participants are eligible for linkage.

### *Survey Data Sources*

Both the NHANES and NHIS are nationally representative cross-sectional surveys of the civilian noninstitutionalized population. NHIS gathers information through a household interview and has a sample size of approximately 39,000 persons each year [1]. From each family in NHIS, one adult aged 18 or older (the "sample adult") is randomly selected to provide information for the sample adult interview. NHANES gathers information through a household interview followed by a physical examination in a mobile examination center, and has a sample size of approximately 5,000 persons each year [2]. Both surveys are important sources of information on the health of the US population. In this analysis the survey data were grouped into earlier (1999-2008) and later (2009-2018) time periods. These groupings ensure that all ESRD estimates are based on unweighted counts of at least 30 ESRD cases, per NCHS standards for presentation of proportions [3].

*End-Stage Renal Disease Administrative Data Sources*

ESRD is a form of permanent kidney failure that requires dialysis or a kidney transplant. ESRD patients who meet certain qualifications are eligible for Medicare, regardless of age [4], and Medicare has provided insurance coverage for approximately 80% of people with ESRD in recent years [5]. Although fewer than 1% of Medicare beneficiaries have ESRD, the condition accounts for approximately 7% of Medicare spending [6]. The USRDS is a national surveillance system that collects information about ESRD in the US for patients of all ages and insurance coverage types [7]. In this analysis, information from the 2020 USRDS Annual Report [8] is considered the true value in comparisons to estimates from the linked survey data.

*Data Linkage Eligibility*

Only NCHS survey participants who have provided consent as well as the necessary personally identifiable information (PII) are eligible for linkage to administrative data sources. NCHS survey participants are informed of NCHS' intent to conduct data linkage activities through a variety of informed consent procedures during survey administration, including advance letters, participant brochures, signed consent forms, and questionnaires.

The percentage of survey participants eligible for the ESRD linkage has varied over time and by survey. Between 1999 and 2008, 79% of adult NHANES participants were eligible for linkage; and 44% of sample adult NHIS participants were eligible for linkage. Between 2009 and 2018, 71% of adult NHANES participants were eligible for linkage; and 61% of sample adult NHIS participants were eligible for linkage. Detailed ESRD linkage eligibility rates are available online [9]. The linkage eligibility rates for ESRD differ from the rates for linkages to other administrative data, like the National Death Index [10], because of the distinct methods that were utilized for each linkage. Approval for the ESRD linkage was provided by the NCHS Research Ethics Review Board (ERB) and the linkage was performed only for eligible NCHS survey participants. The NCHS ERB, functions as an Institutional Review Board or IRB, and is an administrative body of scientists and non-scientists that is established to protect the rights and welfare of human research subjects.

*Data Linkage Methodology*

The NCHS-USRDS ESRD data linkage was performed at the person level using PII from eligible adult survey participants and deterministic (rules-based) techniques [11]. In the first phase of the linkage, USRDS attempted to match survey participants to ESRD records using Social Security Number (SSN), Medicare Health Insurance Claim Number (HICN, when available), or full name, date of birth, and sex. In the second phase, NCHS evaluated each preliminary link to determine whether it included enough matching fields to be considered a true link. Examples of final matches include survey records with the same SSN, month of birth, year of birth and sex as a patient record in the USRDS database.

## Methods

*Data Quality Assessment*

The Federal Committee on Statistical Methodology (FCSM) framework for data quality informed this assessment [12]. The framework divides data quality into three domains, which are utility, objectivity, and integrity, and within each domain, there are 2 to 5 dimensions. An assessment of each dimension of data quality is beyond the scope of this project. Instead, this analysis focused on the domain of objectivity and the dimension of accuracy and reliability. Accuracy is defined as the closeness of an estimate from a data product to its true value whereas reliability refers to the consistency of results when the same phenomenon is estimated more than once under similar conditions. To conduct this data quality assessment, we compared ESRD prevalence estimates based on the linked survey data and the USRDS Annual Report. The following criteria were used to evaluate similarities and differences in estimates: 1. If the prevalence reported in the Annual Report fell within the confidence interval of the survey estimate, the values were described as "similar." 2. If the prevalence reported in the Annual Report fell outside the confidence interval of the survey estimate, the values were described as "different (higher/lower)."

*Weight Adjustment*

As described above, some survey participants are not eligible for record linkage, which creates the potential for bias similar to nonresponse bias. To address this, survey weights were adjusted for linkage eligibility. The weight adjustment model used a three-way interaction between sex, age group and race/ethnicity [13], which are the same demographic factors that were used in post-stratification of the original sample weights. Analyses used survey sample weights and Taylor series linearization to estimate standard errors, accounting for complex survey design.

*Prevalence of ESRD from Survey Data*

Prevalent ESRD cases from the linked survey data were defined as ESRD diagnosis prior to participation in the NCHS survey. Cases were identified using diagnosis date from the administrative data and survey interview date. Prevalence of ESRD was calculated for both pooled survey year groups, 1999-2008 and 2009-2018, using adjusted survey weights. The weighted percentage of US adults with ESRD along with Korn and Graubard confidence intervals was estimated for both NHANES and NHIS. Results were then converted to cases per 1,000,000 for interpretability. The results reflect ESRD case counts among civilian, non-institutionalized US adults.

*Prevalence of ESRD from Administrative Data*

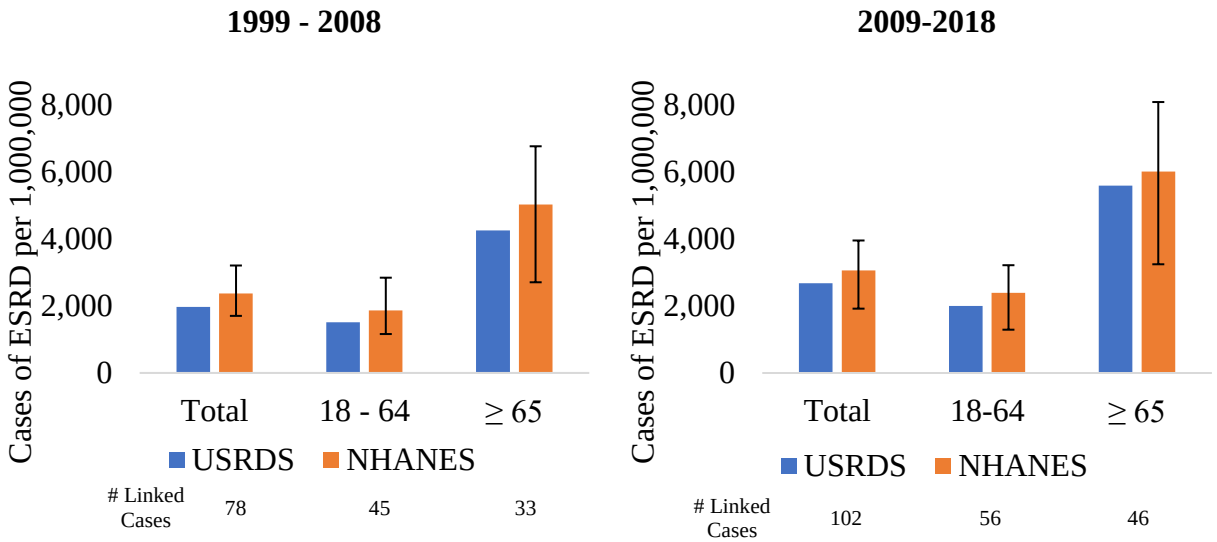
The USRDS Annual Report was used to determine the reported prevalence of ESRD nationwide, which was considered the true value in this assessment [14]. Annual ESRD prevalence estimates were pooled to align with the pooled survey data, and proportions were converted to cases per 1,000,000 for interpretability. The USRDS prevalence estimates excluded records for some ESRD patients missing certain demographic data.

## Results

ESRD prevalence estimates based on the linked NHANES data were similar to the true values reported in the USRDS Annual Report for both time periods included in this analysis (Figure 1). The survey estimates were similar to the true values for all groups, and the true values fell well within the confidence intervals of the survey estimates. In analyses of pooled 1999 – 2008 data, the estimated prevalence among all adults was 2,371 per million population (pmp) (95% CI = 1,705 – 3,210), and the prevalence calculated from the USRDS Annual Report was 1,968 pmp. Among adults aged 18 – 64, the estimated prevalence was 1,869 pmp (95% CI = 1,162 – 2,846), compared with 1,512 pmp from the USRDS Annual Report. Among adults aged 65 and over, the estimated prevalence was 5,034 pmp (95% CI = 3,296 – 7,359), compared with 4,257 pmp from the USRDS Annual Report.

In analyses of pooled 2009 – 2018 data, the estimated prevalence among all adults was 3,062 pmp (95% CI = 2,166 – 4,202), and the prevalence calculated from the USRDS Annual Report was 2,678 pmp. Among adults aged 18 – 64, the estimated prevalence was 2,395 pmp (95% CI = 1,571 – 3,497), compared with 2,006 pmp from the USRDS Annual Report. Among adults aged 65 and over, the estimated prevalence was 6,019 pmp (95% CI = 3,945 – 8,791), compared with 5,597 pmp from the USRDS Annual Report. The confidence intervals were wide due to the modest sample sizes of the linked NHANES data and the small number of ESRD cases (unweighted) in the linked data (78 in 1999 –2008 and 102 in 2009–2018). However, all survey estimates met the NCHS Data Presentation Standards for Proportions [3].

**Figure 1. End-Stage Renal Disease Prevalence Reported in the United States Renal Data System Annual Report, and Estimated Using Linked National Health and Nutrition Examination Survey Data, Adults Aged 18 or Older, 1999 – 2008 and 2009 - 2018**

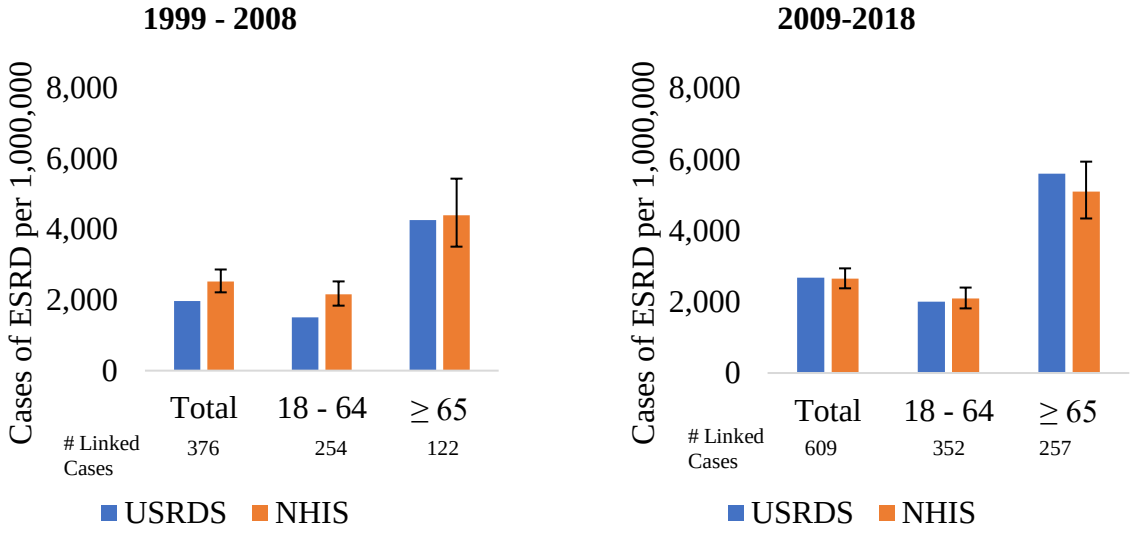



SOURCE: United States Renal Data System Annual Data Report; and National Center for Health Statistics, National Health and Nutrition Examination Survey-United States Renal Data System End-Stage Renal Disease Linked Files.

The accuracy of ESRD prevalence estimates generated using the linked NHIS data varied by time period and age group (Figure 2). In analyses of pooled 1999 – 2008 data, survey estimates were higher than the true value for the total study population and adults aged 18 - 64, and the true value fell outside the confidence intervals for these estimates. Survey estimates were similar to the true values for adults aged 65 and over, and the true value fell within the confidence interval. The estimated ESRD prevalence among all adults was 2,523 pmp (95% CI = 2,215 – 2,862), and the rate calculated from the USRDS Annual Report was 1,968 pmp. Among adults aged 18 – 64, the estimated prevalence was 2,162 pmp (95% CI = 1,838 – 2,526), compared with 1,512 pmp from the USRDS Annual Report. Among adults aged 65 and over, the estimated prevalence was 4,394 pmp (95% CI = 3,508 – 5,434), compared with 4,257 pmp from the USRDS Annual Report.

In analyses of pooled 2009 – 2018 data, survey estimates were similar to the true values for all groups, and the true values fell within the confidence intervals. The estimated prevalence among all adults was 2,648 pmp (95% CI = 2,379 – 2,939), and the rate calculated from the USRDS Annual Report was 2,678 pmp. Among adults aged 18 – 64, the estimated prevalence was 2,092 pmp (95% CI = 1,815 – 2,399) compared with 2,006 pmp from the USRDS Annual Report. Among adults aged 65 and over, the estimated prevalence was 5,092 pmp (95% CI = 4,340 – 5,936), compared with 5,597 pmp from the USRDS Annual Report. The confidence intervals were likely narrower than those from the NHANES due to the larger sample sizes in the linked NHIS and the larger number of ESRD cases (unweighted, 376 in 1999 –2008 and 609 in 2009–2018) compared to the linked NHANES and ESRD data.

**Figure 2. End-Stage Renal Disease Cases Reported in the United States Renal Data System Annual Report, and Estimated Using Linked National Health Interview Survey Data, Adults Aged 18 or Older, 1999 – 2008 and 2009 - 2018**




SOURCE: United States Renal Data System Annual Data Report; and National Center for Health Statistics, National Health Interview Survey-United States Renal Data System End-Stage Renal Disease Linked Files.

## Conclusions

Estimates from linked NHANES-USRDS ESRD data were similar to USRDS ESRD benchmarks, but there were some differences between benchmarks and NHIS-USRDS ESRD data. In earlier NHIS data, when linkage eligibility was low, confidence intervals do not overlap with benchmarks for the study population overall and for adults aged 18-64. Survey estimates are similar to benchmarks and confidence intervals do overlap in later NHIS data, when linkage eligibility was higher. In the NHANES linked data, linkage eligibility was high in both time periods, and survey estimates were similar to benchmarks for all age groups.

## Discussion

This analysis demonstrates the importance of benchmark comparisons in assessments of data quality. However, it must also be emphasized that benchmark comparisons are just one dimension of assessing data quality using the FCSM framework. In this analysis the exclusion of cases with missing demographic information from the benchmark data source might account for some differences in estimates. The findings presented here are consistent with previous research which has found lower linkage eligibility rates coincide with higher "linkage eligibility bias" [13].

## References


1. NCHS. *About the National Health Interview Survey*. 2022 9/20/2022]; Available from: https://www.cdc.gov/nchs/nhis/about_nhis.htm.
2. NCHS. *About the National Health and Nutrition Examination Survey*. 2017 9/20/2022]; Available from: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm.
3. Parker, J.D., et al., *National Center for Health Statistics Data Presentation Standards for Proportions.* Vital Health Stat 2, 2017(175): p. 1-22.
4. CMS. *End-Stage Renal Disease (ESRD)*. [cited 2022 January 5th]; Available from: https://www.medicare.gov/basics/end-stage-renal-disease.
5. USRDS. *Healthcare Expenditures for Persons with ESRD*. 2022 [cited 2022 December 5th]; Available from: https://usrds-adr.niddk.nih.gov/2022/end-stage-renal-disease/9-healthcare-expenditures-for-persons-with-esrd.
6. USRDS. *2018 USRDS Annual Data Report*. 2018 9/20/2022]; 519]. Available from: https://www.usrds.org/media/1734/v2_c09_esrd_costs_18_usrds.pdf.
7. NIH. *About USRDS*. [cited 2022 December 5th]; Available from: https://www.niddk.nih.gov/about-niddk/strategic-plans-reports/usrds/about-usrds#are-only-medicare-esrd-patients.
8. USRDS. *2021 USRDS Annual Data Report*. 2021 10/31/2022]; Available from: https://adr.usrds.org/2021.
9. NHCS. *Linked NCHS-USRDS ESRD - Sample Sizes and Unweighted Percentages by Survey*. 2021 9/20/2022]; Available from: https://www.cdc.gov/nchs/data/datalinkage/MatchRate-Tables.pdf.
10. NCHS. *NCHS Data Linked to NDI Mortality Files*. 2022; Available from: https://www.cdc.gov/nchs/data-linkage/mortality.htm.

11. NCHS. *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*. 2021; Available from: https://www.cdc.gov/nchs/data/datalinkage/NCHS-ESRD-Methodology-Report.pdf.
12. Federal Committee on Statistical Methodology *A Framework for Data Quality*. 2020 September 2020; Available from: https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf.
13. Aram, J., et al., *Assessing Linkage Eligibility Bias in the National Health Interview Survey.* Vital Health Stat 2, 2021(186): p. 1-28.
14. USRDS, *2020 USRDS Annual Data Report*. 2020.