Synthetic Linked 2018 NHIS-HUD-CMS Data: User Guide

Data Release Date: May 15, 2025

Document Version Date: May 15, 2025

Division of Analysis and Epidemiology
National Center for Health Statistics
Centers for Disease Control and Prevention
datalinkage@cdc.gov

Suggested Citation: National Center for Health Statistics. Division of Analysis and Epidemiology. Synthetic Linked 2018 NHIS-HUD-CMS Data: User Guide, May 2025. Hyattsville, Maryland. Available at the following address: https://www.cdc.gov/nchs/data-linkage/synthetic-linked-2018-nhis-hud-cms-data-user-documentation.pdf

Table of Contents

The National Center for Health Statistics (NCHS) Data Linkage Program is designed to enhance the scientific value of NCHS's population-based surveys to support public health surveillance, evidence-based policymaking, and patient-centered outcomes research. The NCHS Data Linkage Program has worked with the Georgia Tech Research Institute to develop a pilot project to produce publicly available synthetic linked data files that appropriately balance the data needs of our wide-ranging user community with our obligations to protect the confidentiality of NCHS survey participants. This document provides information on the creation of the synthetic linked data, evaluation of the utility of the synthetic linked data, and a description of the process created to enable users of the synthetic linked data to verify regression model results against the original restricted-use linked data. The NCHS Disclosure Review Board approved the release of these synthetic linked files. NCHS welcomes feedback and suggestions on how the synthetic linked datasets and/or the verification process can be improved for future synthetic linked data releases. Please send all feedback to datalinkage@cdc.gov.

1. INTRODUCTION

The NCHS Data Linkage Program links data collected by NCHS's population-based and provider-based surveys with health-related administrative data sources to create new data resources that can support robust research and policy evaluation studies by enabling research into the factors that influence disability, chronic disease, health care utilization, morbidity, and mortality.

Due to confidentiality concerns, NCHS linked data files mainly are available as restricted-use files that must be accessed by external researchers through NCHS or Federal Statistical Research Data Centers (RDC). To increase access to these novel data sources, NCHS has engaged in a pilot project to create public-use synthetic linked data files containing linked NCHS survey data and administrative data. Synthetic data can be used to develop public-use files from confidential or restricted data by estimating a statistical model for the joint distribution of the confidential data and generating simulated values from the model as public-use files. This can reduce disclosure risks since the synthetic records do not correspond to actual individuals in the confidential data. Synthetic data are increasingly being used to reduce access barriers to restricted data while maintaining disclosure protections [1]. In addition, NCHS has created a verification process where users can compare selected model results from the synthetic linked data with results based on the original linked data files.

The 2018 National Health Interview Survey (NHIS) was selected for inclusion in the synthetic linked data pilot project as it represented the most current NCHS data that had been linked to health-related administrative data sources when the project was initiated. The 2018 NHIS was previously linked to multiple health-related administrative data sources including Medicare enrollment, utilization, and expenditure data from the Centers for Medicare & Medicaid Services (CMS) and federal housing assistance program participation data from the U.S. Department of Housing and Urban Development (HUD).

This document describes the public-use Synthetic Linked 2018 NHIS-HUD-CMS data, which includes selected variables from the 2018 NHIS public-use data files, the linked 2018 NHIS-HUD data files, the linked 2018 NHIS-CMS Medicare data files, as well as contextual information based on NHIS survey participant Zip Code Tabulation Areas (ZCTAs) from the Agency for Healthcare Research and Quality (AHRQ)'s Social Determinants of Health (SDOH) database.

2. OVERVIEW OF THE SYNTHETIC LINKED DATA

The purpose of the Synthetic Linked 2018 NHIS-HUD-CMS data is to provide access to linked data that are usually accessed only within a secured research environment due to confidentiality concerns. The Synthetic Linked 2018 NHIS-HUD-CMS data consists of 25 synthetic data files, called implicates, which are intended to be analyzed together (see ANALYSIS OF SYNTHETIC IMPLICATES). The synthetic linked data integrates person-level micro-data from the following data sources:

- The public-use 2018 NHIS data files
- The restricted-use 2018 NHIS-HUD linked data files
- The restricted-use 2018 NHIS-CMS Medicare linked data files
- ZCTA-level variables from the AHRQ SDOH database, 2018 release

The population represented in the Synthetic Linked 2018 NHIS-HUD-CMS data is limited to 2018 NHIS survey participants 18 years or older who were linkage eligible for the NHIS-HUD data linkage (i.e., provided consent to link and necessary personally identifiable information for linkage). Variables from all data sources were subject to recoding to create synthetic data variables. A complete list of variables included in the Synthetic Linked 2018 NHIS-HUD-CMS data are available here: Codebook for Synthetic Linked 2018 NHIS-HUD-CMS Data

Background on Source Data Files

Public-use 2018 National Health Interview Survey

NHIS is a nationally representative, cross-sectional household interview survey that serves as an important source of information on the health of the civilian, noninstitutionalized population of the United States. The sample design is a probability design that permits the representative sampling of households and noninstitutional group quarters [2]. For the 2018 survey design, from each family within a sampled household, one “sample adult” aged 18 years or older was randomly selected to complete a detailed questionnaire. Only sample adults were included in the Synthetic Linked 2018 NHIS-HUD-CMS data.

- Eighteen variables derived from the 2018 NHIS public-use data files (person, sample adult, imputed income, and family files) were included. For detailed information on the NHIS’s content and methods, refer to the 2018 NHIS Survey Description document.

Restricted-use 2018 NHIS linked to HUD Housing Assistance Program Data

HUD is the federal agency responsible for overseeing domestic housing programs and policies. While HUD administers various housing and community development programs, the linkage with 2018 NHIS focuses on HUD’s three largest housing assistance programs: Housing Choice Vouchers (HCV), Public Housing (PH), and Multifamily (MF) programs. People receiving housing assistance from HUD 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. For more information on HUD programs, their administration, and HUD data systems, please refer to A Primer on HUD Programs and Associated Administrative Data.

NCHS previously linked eligible survey participant information collected as part of the 2018 NHIS to administrative records from HUD. More detailed information describing the methods used to conduct the NHIS-HUD linkage is available in The Linkage of the National Center for Health Statistics (NCHS) Survey Data to U.S. Department of Housing and Urban Development (HUD) Administrative Data: Linkage Methodology and Analytic Considerations.

- Five variables derived from the NHIS-HUD linked data were included. For detailed information on the NCHS-HUD linked data, refer to NCHS Data Linked to HUD Housing Assistance Program Files.
- Variables from the linked NHIS-HUD data are available for adults aged 18 years or older.

Restricted-use 2018 NHIS linked to CMS Medicare Data

Medicare is the federal health insurance program for people aged 65 years or older, people under age 65 with qualifying disabilities, and people of all ages with end-stage renal disease. For more information on the Medicare program, please refer to the CMS website.

NCHS previously linked eligible survey participant information collected as part of the 2018 NHIS to Medicare claims and enrollment data from CMS. More detailed information describing the methods used to conduct the NHIS-CMS linkage is available in The Linkage of National Center for Health Statistics Surveys to Medicare Enrollment, Claims/Encounters and Assessment Data (2014-2018): Methodology and Analytic Considerations.

- Eight variables derived from the NHIS-CMS Medicare linked data were included. For detailed information on the NCHS-CMS Medicare linked data, refer to NCHS Data Linked to CMS Medicare Data Files.
- Variables from the NHIS-CMS Medicare linked data are available for adults aged 65 years or older.

The population of the Synthetic Linked 2018 NHIS-HUD-CMS data represents NHIS adults aged 18 years or older, who were *linkage eligible for the HUD data linkage*. Synthetic records representing survey participants who were eligible for the CMS linkage but did not link to CMS data will have a value of '0' for the variable CMS_LINKED and will have a missing value for all Medicare variables. Additionally, Medicare variables are included for synthetic records only for NHIS survey participants who were 65 years of age or older at the time of the NHIS interview.

AHRQ Social Determinants of Health Database, 2018 Release

The AHRQ Social Determinants of Health (SDOH) Database was developed to make available SDOH-focused contextual data without the need to access multiple sources. The SDOH Database includes variables derived from the Decennial Census and American Community Survey for ZIP Code Tabulation Areas (ZCTAs). ZCTAs are geographic features made by the U.S. Census Bureau to approximate ZIP Codes. Although most ZIP Codes have a matched ZCTA Code (which have the same five digits as the ZIP Codes they represent), ZCTA boundaries are not exact representations of the United States Postal Service's (USPS) ZIP Code delivery areas. For more information on ZCTAs please refer to ZIP Code Tabulation Areas (ZCTAs).

- Four ZCTA variables derived from AHRQ's SDOH database were included. More information on the AHRQ SDOH database is available at: https://www.ahrq.gov/sdoh/data-analytics/sdoh-data. html.
- ZCTA variables are available for adults aged 18 years or older.

3. CREATION OF THE SYNTHETIC LINKED DATA

Overview

As previously noted, the sample design of the 2018 NHIS is a probability design [2]. To account for the NHIS sampling design, a modification to the original approach to synthesize data described by Rubin [3] was used. First, a synthetic population was created by replicating the observed data according to their NHIS-HUD linkage-eligibility adjusted weights. Second, to reduce file size, a simple random sample was drawn from this synthetic population. Third, using the simple random sample data, synthesis models were estimated and used to create fully synthetic data with the same sample size as the original linked data file. The process was repeated to create 25 synthetic datasets. Survey weights are not included on the synthetic files, as the resulting files can be analyzed as if they were simple random samples [3, 4]. A similar approach is described by Mathur, Reiter, and Si [5]. The three steps are described in more detail below:

The Synthetic Linked 2018 NHIS-HUD-CMS data consists of 25 synthetic data files, called implicates, which are intended to be analyzed together. The use of implicates produces better estimates of the uncertainty of analytic estimates, i.e., proportions and regression coefficients, calculated from the synthetic data. The synthesis process adds uncertainty to the sampling error and other sources of variance already present in the restricted-use linked survey data. Releasing or analyzing

just a single synthetic linked file would not account for this extra uncertainty and thus would understate the uncertainty of estimates from the synthetic linked data.

Step 1: Generating Synthetic Population

To create the synthetic populations, the SimPop package in R was used to implement a bootstrapping approach. Records in the original data file were partitioned into unique strata and primary sampling unit (PSU) combinations for age, sex, and race and Hispanic origin. The NHIS-HUD linkage eligibility weights were summed for the records within each unique NHIS strata-PSU combination to get the number of individuals, *N*. Within each strata-PSU combination, records were sampled with replacement with probability equal to the weight divided by *N*.

For example, suppose that individual i in strata-PSU j has a weight of w_i=250 and that the sum of the weights of all individuals in strata-PSU j is $\widehat{N}_j$=5,000. Then, individual i is sampled with probability 0.05 (250/5,000) and appears in the synthetic population approximately 250 times. The resulting synthetic population has approximately 249 million records, which is consistent with the weighted population of the 2018 NHIS-HUD original data when using the linkage-eligibility adjusted weights.

Step 2: Reducing Synthetic File Size

In the second stage, a simple random sample of 25,000 individuals without replacement from the synthetic population is drawn. This sample size was determined based on the largest number of records that could be drawn within the NCHS secured computing environment without running into memory issues within the synthesis using the synthpop package in R.

Step 3: Estimating Synthetic Models

The random samples selected in step 2 cannot be publicly released as synthetic data files since they correspond to actual individuals. Therefore, to estimate the synthetic models, draws from synthesis models replace variable values, with the exception for the key variables and ZCTA variables.

Sequential Classification and Regression Tree (CART) models [6] were implemented using the synthpop package. Specifically, for each record, the actual values for the NHIS key variables were used for the CART models: age, sex, and race and Hispanic origin. Also, the actual values of the ZCTA level variables were maintained. The modeling also incorporated data rules, such as no pregnant males (see Constraints on Synthesis).

The variables were synthesized in the following sequence: HUD_LINKED, CMS_LINKED, HUD_ASSISTCONCUR, HUD_ASSIST2YPRIOR, HUD_ASSIST5YPRIOR,HUD_ASSISTAFTER, CMS_DUALENROLL, CMS_FF_MONTHS, CMS_MA_MONTHS, CMS_IPVISITS,CMS_EDVISITS, CMS_TOTALPAYMENTS, CMS_VITALSTATUS, NHIS_INSURANCE, NHIS_MARITAL, NHIS_HOUSEOWN, NHIS_PLACEFORCARE, NHIS_DISABILITY, NHIS_EMPLOYMENT, NHIS_EDUCATION, NHIS_SOCIALASSIST, NHIS_FLUVAX, NHIS_PSYCHDISTRESS, NHIS_CHRONICCOND, NHIS_HEALTH, NHIS_POVERTY, NHIS_PREGNANT, NHIS_SMOKING. Descriptions of each of the variables are provided in Codebook for Synthetic Linked 2018 NHIS-HUD-CMS Data.

After the modeling, 22,426 records were randomly drawn, which corresponds to the number of sample adults eligible for the 2018 NHIS-HUD linkage (n=22,472) minus 46 records that were excluded due to potentially problematic values, such as outliers, in the synthetic data. These observations comprise one of the synthetic implicates. The process was then repeated, starting with a new random sample of 25,000 individuals from the synthetic population) to create the desired number of implicates. For the Synthetic Linked 2018 NHIS-HUD-CMS data, 25 implicates were generated as a reasonable number that should be sufficient for reflecting the additional uncertainty added to the data by the synthesis process.

ZCTA-Level AHRQ SDOH Variables

The four ZCTA measures included in the Synthetic Linked 2018 NHIS-HUD-CMS data include information on the ZCTA population's income level, health insurance coverage, internet access, and percentage of the population younger than 65 years of age with Medicaid. The ZCTA level variables are appended to the original NHIS records, prior to data synthesis, by matching the ZCTA codes in the restricted NHIS-linked file.

In the 2018 release of the AHRQ SDOH database, there are 32,989 individual ZCTA codes for the 50 US states and DC, with missing values for some of the ZCTA codes. For the four ZCTA variables selected for inclusion on the synthetic linked data file, the number with missing values ranges from 414 to 578. The missing values were imputed using a scikit-learn IterativeImputer with a Random Forest Regressor at each stage. The regression step used 50 estimators, a maximum depth of 20, and bootstrap sampling using half of the available samples. The result of the imputation process for the four variables is presented in the following table. The "original" values were computed using only the non-missing values for each variable.

Variable Description	Original Mean	Imputed Mean	Original Median	Imputed Median
Income level	14.388	14.508	11.829	12.000
Internet Access	75.130	75.103	77.078	77.183
Medicaid Coverage for age < 65	19.045	19.002	16.667	16.791
No Health Insurance Coverage	9.068	9.114	7.143	7.229

After missing values were imputed, the national median value for each of the four variables was calculated. A new categorical variable is created for each ZCTA variable, indicating whether a given ZCTA value is greater than the median or less than or equal to the median. These median-binned ZCTA level categorical variables are used as the synthetic values themselves.

Constraints on Synthesis

The synthesis process could generate records with illogical relationships between variables, even if that pattern is not seen in the original data. Thus, the following rules were put in place to ensure that no invalid records could be synthesized:

- All variables related to HUD program participation were set to MISSING for records representing survey participants not linked to HUD administrative data.
- The pregnancy variable was set to MISSING for all records representing male survey participants.
- The pregnancy variable was set to "Not pregnant" for all records representing female survey participants over the age of 50.
- All variables related to Medicare program enrollment and utilization were set to MISSING for records representing survey participants younger than age 65.
  NOTE: Vital Status and Dual Enrollment (Medicare/Medicaid) status were also set to MISSING for all records representing survey participants younger than age 65 since the information was originally derived from the linked NHIS-CMS data.
- All variables related to Medicare program enrollment and utilization were set to MISSING for records representing survey participants not linked to CMS administrative data.
- The CMS linkage status variable was set to LINKAGE-ELIGIBLE BUT NOT LINKED for records with zero months of Medicare coverage.

Disclosure and Re-Identification Risk Assessment

After the creation of the synthetic data, disclosure risk analyses were conducted that resulted in adjustments, including additional variable recoding and adding noise, to reduce re-identification risk and ensure the confidentiality of survey participants. Although most of the variables in the source data files were recoded prior to synthesis to minimize data disclosure risk due to small cell sizes, e.g., the public-use NHIS variable for age in integer years (AGE_P) was recoded into 5-year age categories prior to synthesis to reduce re-identification risk, during the synthetic data generation process, additional recodes were undertaken to support the efficient and accurate generation of the synthetic data, as well as to reduce re-identification risk. In addition, the coarse binning process applied to the ZCTA-level variables reduces re-identification risk for these variables, since median binning assigns roughly half of the values to each category. The Synthetic Linked 2018 NHIS-HUD-CMS Data was approved for public release by the NCHS Disclosure Review Board (DRB).

4. ASSESSING SYNTHETIC DATA UTILITY

Data utility metrics that quantify how similar the synthetic data are to the original data are important for assessing the quality of the synthetic data. Several data utility metrics were examined for the synthetic linked 2018 NHIS-HUD-CMS data, including comparisons of univariate distributions as well as higher dimensional metrics (two- or three- dimensional joint distributions) produced using a CART-based generator. It was not feasible to assess all possible relationships between variables in the synthetic and original data. However, results for the relationships examined demonstrated that the synthesis process preserved those relationships in the synthetic data.

In addition to comparing univariate and multivariate joint distributions between the original and synthetic data, additional metrics including the Kullback-Leibler (KL) Divergence (or Relative Entropy Metric) and Total Variation Distance (TVD) Metric were assessed using selected variables. Results from these metrics, particularly KL Divergence metric, were used to inform the decision on the appropriate number of implicates to release. Additionally, since TVD metric tests all possible probability density functions in each dimension (i.e. all crosstabs), the results indicated that there is no variable, pair of variables, or triplet of variables for which the synthetic probability density function failed to closely approximate the original probability density function for the same variables.

Another informative measure of utility is how well synthetic data perform in relation to the original data for a specific type of statistical analysis. Since regression analysis is often applied to NHIS data, randomized regression analysis was used to further examine utility. A randomized set of 100 models were created with the number of explanatory variables ranging from one to four. Due to the categorical nature of the variables, a multinomial logistic regression on both the original data and the synthetic implicates was performed for each of the models with the reference level as the first level of the variable. A survey-based generalized linear model from the svyVGAM package in R was used to estimate the coefficients for the original data. A generalized linear model from the same library was used for each of the synthetic implicates. To combine the results of the implicates, the mitools package in R was used to combine the estimates for direct comparison to the original data. It is important to note that 100 models were originally created but reduced to 43 due to full rank or sparsity issues within the synthetic implicates. These issues were an expected artifact of the randomized models which could produce nonsensical model specifications.

To measure the performance of the synthetic data for selected regression analyses, four metrics were captured for each model parameter:

- Does the sign of the coefficient estimate match for the original and synthetic estimates?
- Is the p-value in the same direction, above or below, with reference to 0.05 for both coefficient estimates?
- Is the synthetic coefficient estimate contained within the confidence interval of the original coefficient estimate?
- What is the overlap of the two confidence intervals?[7]

In the 43 models, 93.8% of parameters had the same coefficient sign. Also, 87.6% of parameters had the p-value in the same direction while 95.6% of the synthetic estimates were contained within the original estimate's confidence interval. Finally, the average confidence interval overlap was found to be 79.1%.

Overall, these regression results were satisfactory in terms of their adherence to the original model estimates. There is some observed variation which is likely due to the sample size used to create the synthetic implicates.

5. ANALYSIS USING SYNTHETIC IMPLICATES

The Synthetic Linked 2018 NHIS-HUD-CMS data are intended to increase researcher access to information from the restricted-use files of the 2018 NHIS linked to administrative data from HUD and CMS Medicare. Researchers should not use the synthetic linked data to analyze variables that were sourced only from the public-use NHIS files, as the NHIS public use files are available for that purpose.

The Synthetic Linked 2018 NHIS-HUD-CMS data include 25 implicate files. Multiple implicates increase the utility of the data by providing more appropriate variance estimates that better reflect the additional uncertainty added to the data by the synthesis process. Analysis using the Synthetic Linked 2018 NHIS-HUD-CMS data are intended to incorporate all 25 synthetic implicate files.

The concept of releasing multiple synthetic implicates is an extension of methods for multiple imputation of missing data proposed by Rubin [8]. Synthetic data researchers have developed alternative combining formulae for fully synthetic and partially-synthetic datasets [9, 10, 11]. Multiple imputation for missing data combining rules proposed by Rubin [8] can be implemented in several statistical software programs. Empirical simulations suggest that Rubin's combining rules for missing data [8] provide a conservative estimate of the variance of estimates from the Synthetic Linked 2018 NHIS-HUD-CMS files.

To assist data users with analysis of the Synthetic Linked 2018 NHIS-HUD-CMS data, sample code for data setup, combining the multiple implicates using Rubin's combining rules, and conducting basic frequency and logistic regression analyses are provided on the NCHS Data Linkage Program website. Sample code is provided for R, SAS, and SAS-callable SUDAAN statistical software packages.

6. VERIFICATION PROCESS

To enhance the utility of the Synthetic Linked 2018 NHIS-HUD-CMS data, users will be able to verify selected results without having to access the original restricted-use linked data. This section provides an overview of the verification process and example output from the verification request process.

Overview of Verification Process

The verification process is currently limited to binomial logistic regression models which will be run in R. The regression on the original restricted-use linked data is a survey design-based estimate using the svyglm function from the survey package, which incorporates the 2018 NHIS-HUD linkage-eligibility adjusted survey weights and the NHIS survey design variables (pseudo-stratum and pseudo-PSU) for variance estimation. Regression models using the Synthetic Linked 2018 NHIS-HUD-CMS data are run using the glm function and are combined into a final synthetic estimate via Rubin's rules [8] using the pool function from the mice package.

The verification process will provide the user with the following four metrics for each estimated model coefficient:

1. Does the sign of the coefficient estimate match for the original and synthetic estimates? (true or false)
2. Is the p-value in the same direction, above or below, with reference to 0.05 for both coefficient estimates? (true or false) Note: The use of the .05 cutoff is provided to facilitate a verification measure.
3. Is the synthetic coefficient point estimate contained within the confidence interval of the original coefficient estimate? (true or false)

4. What is the overlap of the two confidence intervals (original and synthetic)?[7] This is a calculated value between 0-1 and presented to the user in ranges: 0, (0-0.5), [0.5-0.6), [0.6-0.7), [0.7-0.8), [0.8-0.9), [0.9-1.0].

Users can submit a request for verification to datalinkage@cdc.gov using a Template for Requesting Verification Metrics for the Synthetic Linked NHIS-HUD-CMS Data available on the NCHS Data Linkage Program website. Users may select an outcome (dependent) variable and one or more predictor (independent) variables from a prepopulated list. Additionally, the user can specify the reference levels for their predictor/independent variables (by default, the lowest value will be used as the reference level). Lastly, a universe or subpopulation for the model can be selected from the following options:

- All adults (no exclusions)
- Adults aged 65 and over who are linked to Medicare (UNIV_65P_MCARE=1)
- Adults aged 65 and over who are linked to Medicare and had Traditional / Fee-for-Service coverage for at least one month (UNIV_MCARE_FFS=1)
- Adults who linked to one or more years of HUD administrative data (HUD_LINKED=1)
- Adults with family income-to-poverty ratio less than 2 (UNIV_LT2XPOV=1; only available when HUD_ASSISTCONCUR is selected as the outcome variable)
- Adults with family income-to-poverty ratio less than 2 AND housing status indicates rental (UNIV_RENTLT2XPOV=1; only available when HUD_ASSISTCONCUR is selected as the outcome variable)

The verification process is not intended to inform model selection. Rather, the verification measures indicate the similarity in the synthetic and confidential data results and that close similarity does not imply that the confidential data model accurately describes the data, nor do differences imply that the confidential data model fails to describe the data accurately. Please note that interaction terms between two or more variables are not currently supported. Due to finite staffing resources, there may be a limit on the number of verification requests per data user.

Verification Output

Figures 1 and 2, below, are example visuals for the verification output. ***The examples shown are for illustration purposes. They are based on fabricated data and do not use the original restricted-use or Synthetic Linked 2018 NHIS-HUD-CMS data.***

In this example, the model has reported receipt of a flu vaccine in the past year (NHIS_FLUVAX) as the response variable and age group (NHIS_AGE_R2), race and ethnicity (NHIS_RACE_ETH), and respondent's report of a usual place to receive medical care (NHIS_PLACEFORCARE) as the predictors. The reference groups for the predictors are age 18-44, Non-Hispanic White, and response of 'No' for usual place to receive medical care.

Figure 1 displays the histogram viewer, which provides univariate distributions from the synthetic and original data for NHIS variables included in the requested models. Univariate distributions for variables derived from linked HUD and CMS files will not be provided.

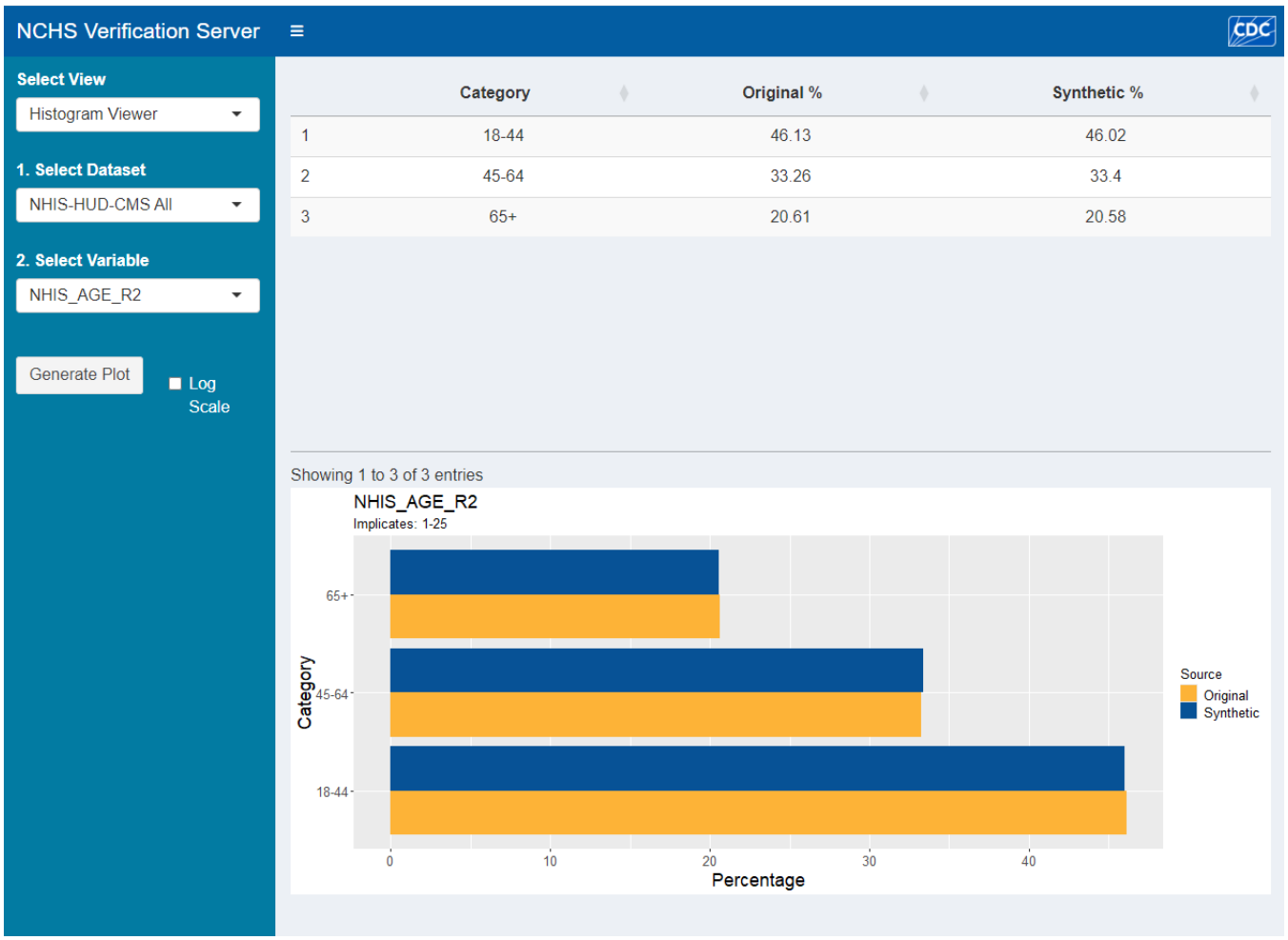

Fig. 1: Comparison of Univariate Distributions

Figure 2 displays the four metrics provided in the verification output. Metrics 1 through 4 specify whether the coefficients in a model run on the synthetic data are similar to those from the same model run on the original restricted-use data.

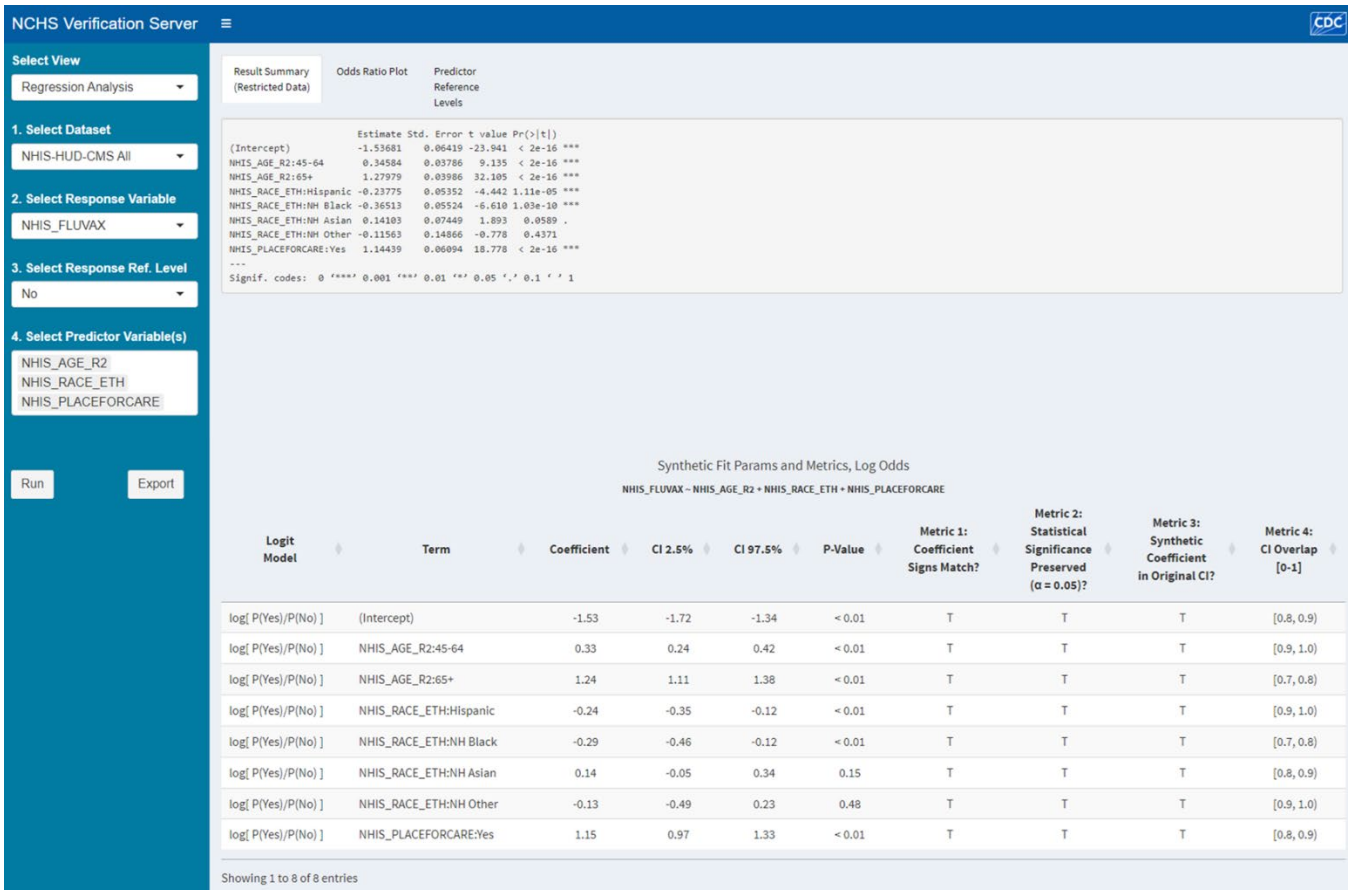

Fig. 2: Verification Metrics for Binomial Logistic Regression Models

For this example, we will interpret the results for NHIS_PLACEFORCARE: Yes:1 *(last row in Figure 2)*.

1. **Does the sign of the coefficient estimate match for the original and synthetic estimates?**

   Metric 1 = T, indicating true. The coefficient signs do match for the original and synthetic estimates.

2. **Is the p-value in the same direction, above or below, with reference to 0.05 for both coefficient estimates?**

   Metric 2 = T, indicating true. Both the original and the synthetic estimate indicate the p-value is in the same direction as less than 0.05.

3. **Is the synthetic coefficient point estimate contained within the confidence interval of the original coefficient estimate?**

   Metric 3 = T, indicating true. The synthetic coefficient estimate (1.15) is within the confidence interval for the original coefficient estimate.

4. **What is the overlap of the two confidence intervals (synthetic and original)?**

   Metric 4 = [0.8, 0.9), indicating that the range of confidence interval overlap for the synthetic data and the original data is from 0.8 up to 0.9.

It is expected that users run their model(s) of interest using the public-use synthetic linked data and ensure that no errors are produced, before submitting a verification request.

Reporting Verification Metrics

Verification metrics provided to data users can be included in publications. We encourage data users to carefully consider the verification metrics when using the synthetic linked data and appropriately report the metrics when publishing results based on the synthetic data. Data users interested conducting additional analyses will be able to access the original restricted-use linked data through the NCHS RDC or the Federal Statistical RDC Network for data users.

REFERENCES

1. Reiter, J.P. (2023). Synthetic Data: A Look Back and A Look Forward. Transactions On Data Privacy 16, 15-24.
2. National Center for Health Statistics. Survey Description, National Health Interview Survey, 2018. Hyattsville, Maryland. 2019.
3. Rubin, D.B. (1993). Statistical Disclosure Limitation. Journal of Official Statistics, 9, 461-468.
4. Reiter, J.P. (2005). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. Journal of the Royal Statistical Society Series A: Statistics in Society, 168(1), 185-205.
5. Mathur, S., Si, Y. and Reiter, J.P. (2024). Fully synthetic data for complex surveys. Survey Methodology, 50(2), 347-373.
6. Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. Journal of Official Statistics, 21, 441 - 462.
7. Karr, A.F., Kohnen, C.N., Oganian, A. Reiter, J.P., and Sanil, A.P. (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician. 60(3), 224-232.
8. Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.
9. Raghunathan, E.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statistics 19, 1-16.
10. Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. Journal of Official Statistics 18, 531544.
11. Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. Survey Methodology 29, 181-189.

ACKNOWLEDGEMENTS

NCHS acknowledges Richard Boyd and Austin Himschoot (GTRI) for the application of data synthesis techniques and Jerome Reiter (Duke University) for providing subject matter expertise in data synthesis and disclosure risk avoidance. The pilot project was conducted with support from the CDC Office of Science and funding from the Office of the Secretary Patient Centered Outcomes Research Trust Fund (OS-PCORTF).