

Synthetic Linked 2016 NHCS-NDI Data: User Guide

Data Release Date: October 8, 2026

Document Version Date: October 8, 2026

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 2016 NHCS-NDI Data: User Guide, October 2026. Hyattsville, Maryland. Available at the following address:
<https://www.cdc.gov/nchs/linked-data/synthetic/index.html>

Table of Contents

1. INTRODUCTION 4

2. OVERVIEW OF THE SYNTHETIC LINKED DATA 5

 Background on Source Data Files 5

3. CREATION OF THE SYNTHETIC LINKED DATA 6

 Overview.....6

 Step 1: Data Partitioning.....6

 Step 2: Estimating Synthetic Models 7

 Step 3: Combining Partitions and Reducing Synthetic File Size 7

 Constraints on Synthesis.....8

 Disclosure and Re-Identification Risk Assessment.....8

4. ASSESSING SYNTHETIC DATA UTILITY 9

5. ANALYSIS USING SYNTHETIC IMPLICATES 9

6. VERIFICATION PROCESS 10

 Overview of Verification Process 10

 Verification Output 10

 Verification Metrics for Anomalous Regressions..... 12

 Reporting Verification Metrics 12

REFERENCES 13

ACKNOWLEDGEMENTS 14

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 2016 National Hospital Care Survey (NHCS) 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 2016 NHCS was previously linked to multiple [health-related administrative data sources](#) including mortality data collected from the National Death Index (NDI).

This document describes the public-use Synthetic Linked 2016 NHCS-NDI data, which includes selected variables from the [2016 NHCS](#), [linked 2016 NHCS – 2016-2017 NDI](#), and [2016 NHCS Enhanced Opioid Identification](#) restricted-use data files.

2. OVERVIEW OF THE SYNTHETIC LINKED DATA

The purpose of the Synthetic Linked 2016 NHCS-NDI 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 2016 NHCS-NDI data consist of 16 synthetic data files, called implicates, which are intended to be analyzed together (see [ANALYSIS OF SYNTHETIC IMPLICATES](#)). The synthetic linked data integrates patient-level micro-data from the following data sources:

- The restricted-use [2016 NHCS data files](#)
- The restricted-use [linked 2016 NHCS – 2016-2017 NDI data files](#)
- The restricted-use [2016 NHCS Enhanced Opioid Identification file](#)

The population represented in the Synthetic Linked 2016 NHCS-NDI data is limited to 2016 NHCS patient records for adults aged 18 and older who were linkage eligible for the NHCS-NDI data linkage (i.e., participating hospitals provided 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 2016 NHCS-NDI data are available here:

[Codebook for Synthetic Linked 2016 NHCS-NDI Data](#)

Background on Source Data Files

Restricted-use 2016 National Health Hospital Survey Data

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

From participating hospitals, NHCS collects data on all IP and ambulatory care visits occurring during the calendar year. In previous years of the survey, hospitals were required to provide data from claims records, but to reduce the burden of reporting on participating hospitals, for the 2016 data collection hospitals were given the option of providing their data in the form of electronic health records (EHR) or as claims records. Thus, participating hospitals provided data in the form of Uniform Bill (UB)-04 administrative claim records or EHR data, where the EHR data are an amalgamation of custom extracts and Consolidated Clinical Documents (CCDs). NHCS collects patient PII (e.g., name, date of birth, and Social Security Number (SSN)), which allows for the linkage of episodes of care across hospital units as well as to other data sources, such as the NDI.

The 2016 NHCS data are available for research use as a restricted-use data file accessible via the NCHS RDC network. The 2016 restricted-use NHCS data file is structured at the encounter level (one or more hospital encounter records) and does not include patient-level variables; patient-level variables were constructed from this data source for use in the synthetic linked NHCS-NDI pilot project.

- 25 variables derived from the restricted-use 2016 NHCS file were included. For detailed information on the NHCS's content and methods, refer to the [NHCS webpage](#).

Restricted-use Linked 2016 NHCS – 2016-2017 NDI Data

Through its data linkage program, NCHS conducted a linkage of the 2016 NHCS with 2016-2017 NDI data. The NDI is a centralized database of U.S. death record information on file in state vital statistics offices. Working with these state offices, NCHS established the NDI as a resource to aid epidemiologists and other health and medical investigators with their mortality ascertainment activities. The records, which are compiled annually, include detailed information on the underlying and multiple causes of death. More detailed information describing the methods used to conduct the NHCS-NDI linkage is available in [The Linkage of the 2016 National Hospital Care Survey to the 2016-2017 National Death Index: Methodology Overview and Analytic Considerations](#).

- 19 variables derived from the restricted-use linked 2016 NHCS-NDI data were included. For detailed information on the linked NHCS-NDI linked data, refer to [NHCS Data Linked to NDI Mortality Data](#).

Restricted-use 2016 NHCS Enhanced Opioid Identification File

The NCHS Division of Health Care Statistics (DHCS) created an [Enhanced Opioid Identification Algorithm](#) to identify opioid-involved hospital encounters in ED and IP settings. The algorithm was designed to capture specific codes reported in hospital encounter records and included natural language processing (NLP) machine learning techniques for scanning physician's clinical notes to identify opioid involvement. An encounter-level file, the 2016 NHCS Enhanced Opioid Identification file was created to identify hospital IP or ED encounter records that met the criteria for opioid involvement.

The restricted-use 2016 NHCS Enhanced Opioid Identification file is available for research use via the NCHS RDC network; however, this data source does not include patient-level variables. Patient-level variables were constructed from this data source for use in the synthetic linked NHCS-NDI pilot project.

- 2 variables derived from the restricted-use 2016 NHCS Enhanced Opioid Identification file were included. For detailed information on the file's creation and methods, refer to [Linked Data on Hospitalizations, Mortality, and Drugs: Data from the National Hospital Care Survey 2016, National Death Index 2016-2017, and the Drug-Involved Mortality 2016-2017](#) documentation.

3. CREATION OF THE SYNTHETIC LINKED DATA

Overview

The Synthetic Linked 2016 NHCS-NDI data file includes patient-level variables constructed from IP or ED hospital encounter records collected in the 2016 NHCS and mortality information for patients who were linked to an NDI record. All 2016 NHCS hospital encounter records include a patient identification number which allowed the hospital encounter data to be summarized at the patient level. The 2016 NHCS is not nationally representative and does not include sample weights. Therefore, sample weights were not used in the synthetic data generation process and analyses conducted on the synthetic linked NHCS-NDI data file should not be considered national estimates of patient data.

The Synthetic Linked 2016 NHCS-NDI data consist of 16 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.

The steps taken to create the synthetic data implicates are described below:

Step 1: Data Partitioning

The first step in the process of creating the synthetic data is the partitioning of the linked input data based on key variables related to the survey design and the characteristics of the linked records to be synthesized. The 2016 NHCS is designed to collect data on patient care from sampled hospitals. The hospital sample design is based on sampling strata defined by hospital type of service (general acute, children's hospitals, psychiatric, and other long term care hospitals). Among the general acute hospitals, sampling strata are further defined by four bed size designations (<50, 50-199, 200-499, and 500+ beds) and four area types (central city in a Metropolitan Statistical Area (MSA) 1 million plus population,

fringe part of MSA of 1 million plus population, MSA of less than 1 million population, and non-MSA). These 2016 NHCS sample design characteristics were used to subset the 2016 NHCS patient-level records linked to the 2016-2017 NDI records into seven separate files (subsets). The seven subsets were synthesized separately due to the potential impact of the sample design characteristics on NHCS variable inter-relationships during the synthetic data generation process. Additionally, four of the seven subsets were subsequently split based on the patient's vital status (alive or deceased) using the variable (MORTSTAT) from the linked 2016 NHCS – 2016-2017 NDI data to ensure adherence of dependent relationships between vital status and other variables in the synthesis process. This resulted in a total of 11 data partitions that were independently synthesized.

Step 2: Estimating Synthetic Models

Sequential Classification and Regression Tree (CART) models [2] were implemented using the R [synthpop](#) package to synthesize the 11 data partitions. The modeling incorporated several data rules described below (see [Constraints on Synthesis](#)) to avoid illogical relationships between variables. Helper variables were created and used during the synthetic data generation process to constrain relationships between variables. The helper variables were only used during the synthesis process and subsequently dropped from the final synthetic data implicates. Despite the use of helper variables, some synthesized records still resulted in illogical variable combinations, and these records were subsequently dropped from the final synthetic file.

For each of the 11 data partitions created in step 1, two variables were sampled in the modeling steps: vital status (MORTSTAT) and the helper variable **OP_LT**. All remaining variables were synthesized in the following sequence: P_AGE_CD, P_RACE_ETHNCTY_CD, P_SEX_CD, P_URBRURAL_CD, MONTH_LAST_ENC, **DISCHARGE_DEAD**, DSCHRG_STUS_CD, LAST_ENC_SETTING, TOT_ED_ENC, TOT_IP_ENC, TOT_OPIOID_ENC, ED_ENCOUNTER_POSTIP_IND, HEART_DISEASE_IND, TOT_ICU_DAYS, **ALZHEIMER**, UCOD_LEADING, TIME_TO_DEATH_CD, **DRUG**, COPD_IND, DIABETES_IND, MENTAL_HEALTH_IND, SEPSIS_IND. Descriptions of each of the variables are provided in [Codebook for Synthetic Linked 2016 NHCS-NDI Data](#).

NOTE: **OP_LT**, **DISCHARGE_DEAD**, **ALZHEIMER**, and **DRUG** are helper variables used to enforce logical constraints on the relationships between variables but are not included on the final synthetic data file.

- **OP_LT** is a binary variable with the values (0) the total number of opioid encounters is equal to the sum of IP and ED encounters and (1) the total number of opioid encounters is less than the sum of IP and ED encounters.
- **DISCHARGE_DEAD** is a binary variable with the values (0) the patient was discharged alive and (1) the patient was discharged deceased.
- **ALZHEIMER** is a binary variable with the values (0) the patient's leading cause of death was not Alzheimer's disease and (1) the patient's leading cause of death was Alzheimer's disease.
- **DRUG** has four categories indicating (0) patient assumed deceased with no indication of either drug overdose death or drug overdose death with opioid-involvement (1) patient assumed deceased with indication of both drug overdose death and drug overdose death with opioid-involvement (2) patient assumed deceased with indication of either (only one or the other) drug overdose death or drug overdose death with opioid-involvement and (3) patient assumed alive or patient assumed deceased, but information related to drug overdose death and drug overdose death with opioid-involvement is not available.

Step 3: Combining Partitions and Reducing Synthetic File Size

After the modeling, records were randomly drawn from each synthetic data partition to equal the original record count of the input partition plus an additional 1% (to account for records to be dropped due to illogical variable combinations). The modeling and sampling are repeated for the subsequent partitions until all 11 partitions are synthesized. The 11 synthetic data partitions are combined with the other data partitions from their respective subsets to create seven synthetic data subsets (that correspond with the subsets created for the input data). Records that do not meet the constraint criteria are dropped from each subset. Additional records are dropped (if necessary) to create synthetic data subsets that match the

total record counts in the input data subsets. Finally, the seven synthetic data subsets are combined to form a single implicate, containing 2,916,833 records, which corresponds to the number of linkage-eligible 2016 NHCS adult patients with non-missing information for age, sex, and patient residence urban/rural status, minus patients that were assumed deceased from non-NDI source (i.e., hospital discharge status) but not successfully linked to the NDI. This process is repeated until the 16 implicates are created. For the Synthetic Linked 2016 NHCS-NDI data, 16 implicates were generated as a reasonable number that should be sufficient for reflecting the additional uncertainty added to the data by the synthesis process.

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:

- Synthesized records must have at least one ED encounter or IP encounter.
- The total number of patient ED encounters could not be set to “none” for records representing patients whose last encounter setting was ED.
- The total number of patient IP encounters could not be set to “none” for records representing patients whose last encounter setting was IP.
- The indicator for ED encounter within 7 days of IP discharge was set to “No” for records representing patients with no IP encounters or no ED encounters.
- The total number of patient ICU days was set to 0 for records representing patients who did not have an IP encounter.
- The total number of opioid-related encounters could not exceed the total number of encounters (represented by total IP encounters plus total ED encounters).
- The Hospital Discharge Status (DSCHRG_STUS_CD) was reset to “Other discharge or discharge status unknown” for records representing patients who were assumed alive, but whose discharge status was initially synthesized as “Deceased”.
- The Time to Death (TIME_TO_DEATH_CD) was set to “0-30 days” for records representing patients with a discharge status of deceased.
- The Cause of Death (UCOD_LEADING) was reset to “All other causes” for any synthesized records representing patients aged 18-44 whose cause of death was initially synthesized as Alzheimer’s disease.
- The Drug Overdose Death (DRUG_OVRDSE_DEATH) and Drug Overdose Death Involving Any Opioid (DRUG_OVRDSE_DEATH_ANYOPIOID) indicators could only be synthesized as “Yes” for synthesized records representing patients with underlying cause of death due to “Accidents” or “All other causes”.
- The Drug Overdose Death (DRUG_OVRDSE_DEATH) and Drug Overdose Death Involving Any Opioid (DRUG_OVRDSE_DEATH_ANYOPIOID) indicators were synthesized together using a helper variable to ensure that records coded as Drug Overdose Death Involving Any Opioid were a subset of records coded as Drug Overdose Death.

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 variable constructed from the restricted-use NHCS data to indicate the total number of emergency department encounters (TOT_ED_ENC) was recoded into four categories (0 - No encounters; 1 - One encounter; 2 - Two encounters; 3 - Three or more encounters) 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. The Synthetic Linked 2016 NHCS-NDI 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 2016 NHCS-NDI 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 NHCS data, randomized regression analysis was used to further examine utility. A randomized set of 2,000 models were created with the number of explanatory variables ranging from two to 17. No evidence of separation or coefficient estimates with a value of NA were reported in any of the models. 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.

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

1. Does the sign of the coefficient estimate 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?
3. Is the synthetic coefficient estimate contained within the confidence interval of the original coefficient estimate?
4. What is the overlap of the two confidence intervals?[3]

In the 2,000 models, 96.2% of parameters had the same coefficient sign. Also, 90.4% of parameters had the p-value in the same direction while 61.4% of the synthetic estimates were contained within the original estimate's confidence interval. Finally, there was at least some degree of confidence interval overlap for 86.5% of the coefficients.

Overall, these regression results were satisfactory in terms of their adherence to the original model estimates. There are some observed variation, particularly related to metrics 3 and 4, which are likely attributed to the narrow confidence intervals due to the large sample size.

5. ANALYSIS USING SYNTHETIC IMPLICATES

The Synthetic Linked 2016 NHCS-NDI data include 16 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 2016 NHCS-NDI data are intended to incorporate all 16 synthetic implicate files.

The concept of releasing multiple synthetic implicates is an extension of methods for multiple imputation of missing data proposed by Rubin [4]. Synthetic data researchers have developed alternative combining formulae for fully synthetic and partially-synthetic datasets [5, 6, 7]. Multiple imputation for missing data combining rules proposed by Rubin [4] can be implemented in several statistical software programs. Empirical simulations suggest that Rubin's combining rules for missing data [4] provide a conservative estimate of the variance of estimates from the Synthetic Linked 2016 NHCS-NDI files.

To assist data users with analysis of the Synthetic Linked 2016 NHCS-NDI 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 Public-use Synthetic Linked Data webpage](#). 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 2016 NHCS-NDI 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 regressions on the original restricted-use linked data and the Synthetic Linked 2016 NHCS-NDI data are run using the `glm` function in base R. The estimates derived from each synthetic data implicate are combined into a final synthetic estimate via Rubin's rules [4] 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)? [3] 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 NHCS-NDI Data](#) available on the [NCHS Public-use Synthetic Linked Data webpage](#). 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)
- Only adult patients assumed to be deceased based on NDI linkage (MORTSTAT=1)

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

Figure 1, below, provides a visual example of the verification output. ***The example shown is for illustration purposes. The results are based on fabricated data and do not use the original restricted-use or Synthetic Linked 2016 NHCS-NDI data.***

Synthetic Linked 2016 NHCS-NDI Data: User Guide

In this example, the model has the underlying cause of death of heart disease (UCOD_HEART_DISEASE) as the response variable and race and ethnicity (P_RACE_ETHNCTY_CD) and total number of emergency department (ED) encounter records (TOT_ED_ENC) as the predictors. The reference groups for the predictors are Non-Hispanic White and no ED encounter records. The universe for the model is adult patients assumed to be deceased based on NDI linkage (UNIV_DISEASED=1).

Figure 1 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.

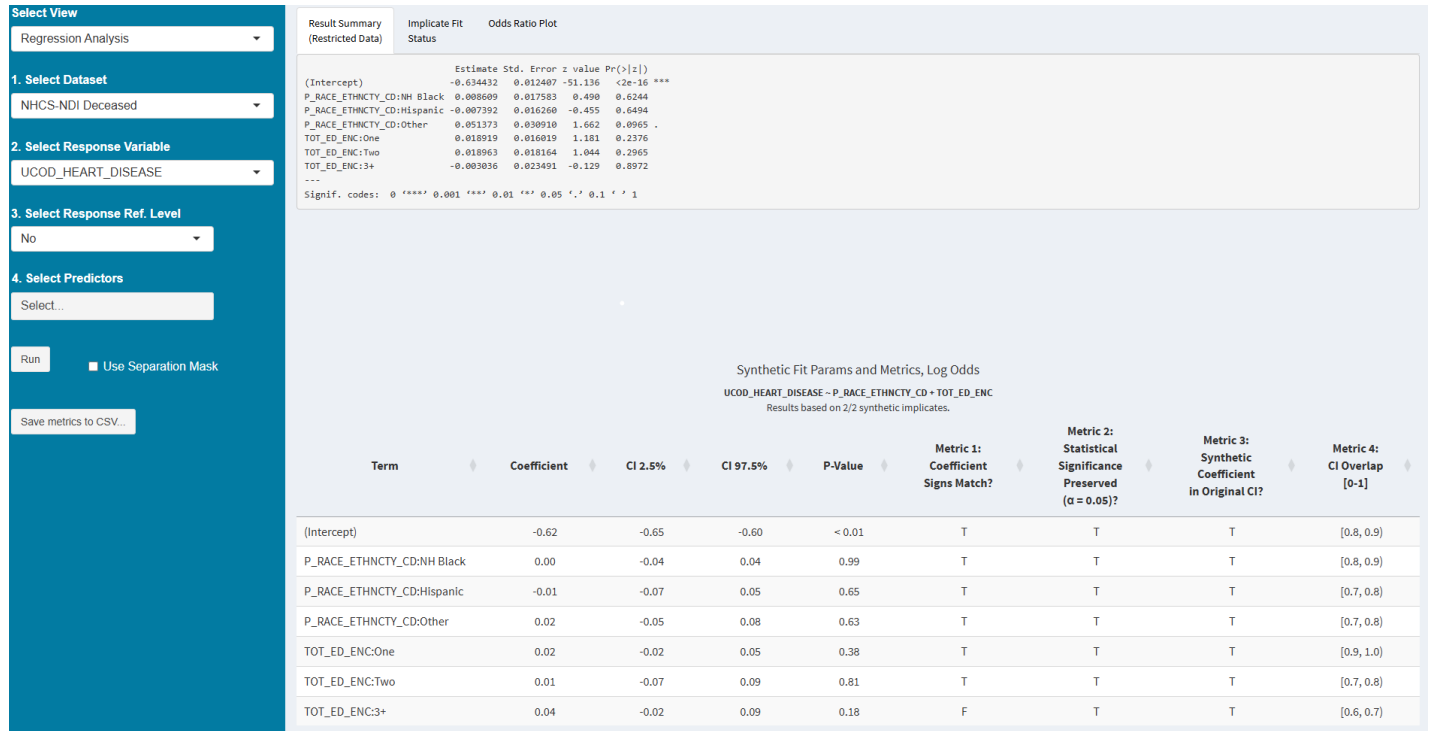


Fig. 1: Verification Metrics for Binomial Logistic Regression Models

For this example, we will interpret the results for TOT_ED_ENC: 3+ (last row in Figure 1).

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

Metric 1 = F, indicating false. The coefficient signs do not 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 (0.04) 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 = [0.6, 0.7), indicating that the range of confidence interval overlap for the synthetic data and the original data is from 0.6 up to 0.7.

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.

Verification Metrics for Anomalous Regressions

The verification metrics can be easily computed if each regression model converges and returns finite values for all coefficients and intercepts. However, some logistic regression models can produce unusable output. The verification process is designed to detect anomalous results, and the following rules are applied before the results are reported to data users:

- Model fit for the synthetic data regression produces an NA for one or more coefficients
 - Coefficients estimated on the synthetic data are returned verbatim
 - No verification metrics are reported for any variable in the model
- Synthetic data regression model estimates all coefficients, but restricted data regression model produces an NA for one or more coefficients. Rules are applied that act as if the coefficient estimate from the confidential data regression is close to zero.
 - Metric 1 (sign match): T, indicating true, is reported when the synthetic data confidence interval includes zero, otherwise F, indicating false, is reported.
 - Metric 2 (p-value): T, indicating true, is reported if the p-value for the synthetic data coefficient exceeds the threshold, otherwise F, indicating false, is reported.
 - Metric 3 (synthetic coefficient inside original CI): F, indicating false, is reported
 - Metric 4 (CI overlap): the interval (0, 0.5) is reported.
- Regressions with coefficients that show evidence of perfect separation – i.e., one or more combinations of the predictor variables can perfectly discriminate the binary outcome - in either the restricted or synthetic data regressions
 - Metric 1 (sign match): reported with no modification
 - Metric 2 (p-value): reported with no modification
 - Metric 3 (synthetic coefficient inside original CI): F, indicating false, is reported
 - Metric 4 (CI overlap): the interval (0, 0.5) is reported.

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 in 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. Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21, 441 - 462.
3. 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.
4. Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
5. Raghunathan, E.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 19, 1-16.
6. Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* 18, 531-544.
7. 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).